Quantile-Optimal Treatment Regimes

Lan Wang, Yu Zhou, Rui Song and Ben Sherwood

Abstract

Finding the optimal treatment regime (or a series of sequential treatment regimes) based on individual characteristics has important applications in areas such as precision medicine, government policies and active labor market interventions. In the current literature, the optimal treatment regime is usually defined as the one that maximizes the average benefit in the potential population. This paper studies a general framework for estimating the quantile-optimal treatment regime, which is of importance in many real-world applications. Given a collection of treatment regimes, we consider robust estimation of the quantile-optimal treatment regime, which does not require the analyst to specify an outcome regression model. We propose an alternative formulation of the estimator as a solution of an optimization problem with an estimated nuisance parameter. This novel representation allows us to investigate the asymptotic theory of the estimated optimal treatment regime using empirical process techniques. We derive theory involving a nonstandard convergence rate and a non-normal limiting distribution. The same nonstandard convergence rate would also occur if the mean optimality criterion is applied, but this has not been studied. Thus, our results fill an important theoretical gap for a general class of policy search methods in the literature. The paper investigates both static and dynamic treatment regimes. In addition, doubly robust estimation and alternative optimality criterion such as that based on Gini's mean difference or weighted quantiles are investigated. Numerical simulations demonstrate the performance of the proposed estimator. A data example from a trial in HIV+ patients is used to illustrate the application.

KEY WORDS: dynamic treatment regime; nonstandard asymptotics; optimal treatment regime; precision medicine; quantile criterion.

[1]Lan Wang is Professor and Yu Zhou is graduate student, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Emails: wangx346@umn.edu and zhou0269@umn.edu. Rui Song is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. Email: rsong@ncsu.edu. Ben Sherwood is Assistant Professor, School of Business, University of Kansas. Email: ben.sherwood@ku.edu. Dr. Sherwood's work was done when he was a graduate student at University of Minnesota. Wang's research is partly supported by NSF DMS-1512267 and DMS-1712706. Song's research is partly supported by NSF DMS-1555244 and NCI P01 CA142538. We thank the Co-Editor Nicholas Jewell, the AE and two anonymous referees for their constructive comments which help us significantly improve the paper. We also thank Dr. Shannon Holloway at North Carolina State University for proofreading the paper and providing many helpful comments.

1 Introduction

A treatment regime can be described as a function from the space of covariates to the set of treatment options. Depending on the application, a treatment can represent a drug, a device, a program, a policy, an intervention or a strategy. The problem of estimating an optimal treatment regime has recently received considerable attention. Medical doctors have long been interested in tailoring a patient's medical treatment according to the individual's unique genetic information, health history, environmental exposure, needs and preferences. Economists are interested in finding the most effective active labor market programs (job search training, computer training, etc.) for an unemployed job seeker (Frölich (2008), Behncke et al. (2009), Staghøj et al. (2010), Wunsch (2013)). In political science, researchers are interested in selecting the best strategies (personal visits, phone calls, mailings, etc.) to increase voter turnout (Gerber and Green (2000), Imai and Ratkovic (2013)).

Existing work on estimating an optimal treatment regime has mainly focused on the mean-optimal treatment regime, which if followed by the whole population would yield the largest average outcome (assuming a larger outcome is preferable). Popular approaches for estimating mean-optimal treatment regimes include model-based methods such as Q-learning (Watkins and Dayan, 1992; Murphy, 2005b; Chakraborty et al., 2010; Moodie and Richardson, 2010; Goldberg and Kosorok, 2012; Song et al., 2015), A-learning (Robins et al., 2000; Murphy, 2003, 2005a), and model-free or policy search methods (Robins and Rotnitzky, 2008; Orellana and Robins, 2010; Zhang et al., 2012a; Zhao et al., 2012, 2015a). Other relevant work includes Robins (2004); Moodie et al. (2007, 2009); Henderson et al. (2010); Cai et al. (2011); Qian and Murphy (2011); Thall et al. (2011); Imai and Ratkovic (2013); Huang et al. (2015); Tao and Wang (2017), among others. We refer to the recent books (Chakraborty and Moodie, 2013; Kosorok and Moodie, 2016) and review articles (Qian et al., 2012; Chakraborty and Murphy, 2014; Laber et al., 2014; Schulte et al., 2014; Wallace and Moodie, 2014) for a more comprehensive list of references. In econometrics, an

independent line of interesting work explored a decision theory framework for estimating statistical treatment rules (Manski, 2004; Dehejia, 2005; Hirano and Porter, 2009; Stoye, 2009; Bhattacharya, 2009; Bhattacharya and Dupas, 2012; Tetenov, 2012).

In a variety of applications, criteria other than the mean (or the average) may be more sensible. When the outcome has a skewed distribution (e.g., survival time of patients), it may be desirable to consider the treatment regime that maximizes the median of the distribution of the potential outcome. Sometimes, the tail of the potential outcome distribution is of direct importance. When evaluating government job training programs to improve earnings, policy makers may ask which program does best to improve earnings on the lower tail. An optimal treatment regime with respect to the tail criterion is even more attractive if the sacrifice is little at the central part of the potential outcome distribution as compared to the mean-optimal treatment regime. A simple numerical example illustrating phenomenon of this nature is given in Section 2. The same numerical example also reveals that the mean-optimal treatment regime may work poorly (or even have detrimental effect) at the tails.

In this paper, we study a general framework for estimating the quantile-optimal treatment regime in both static and dynamic settings, the latter of which involves estimating a sequence of treatment regimes that may vary over time based on a longitudinal study. Given a class of treatment regimes, we consider a robust estimator of the quantile-optimal treatment regime that does not require specifying an outcome regression model. By now, it has been widely recognized (Qian and Murphy, 2011; Zhang et al., 2012a; Zhao et al., 2012; Matsouaka et al., 2014; Zhao et al., 2015b) that a fundamental challenge in estimating the optimal treatment regime is specifying a reliable outcome model, which describes how the treatment and covariates influence the outcome and how they interact with each other. A misspecified outcome model can result in biased estimation of the optimal treatment regime. The difficulty of specifying outcome models is more pronounced when estimating the optimal

dynamic treatment regime using longitudinal data, for which model-based approaches would require specifying a sequence of outcome models, one for each decision point. However, complete nonparametric estimation of optimal treatment regimes suffers from the curse of dimensionality and does not provide easy-to-interpret treatment regimes.

Although some recent work has made important contributions to estimating the optimal treatment regime without an outcome model (Robins and Rotnitzky, 2008; Robins et al., 2000; van der Laan et al., 2005; Orellana and Robins, 2010; Zhang et al., 2012a, 2013; Zhao et al., 2012, 2015b), they have considered only the mean-optimal criterion and have not studied the asymptotic distribution of the estimated optimal treatment regime. In fact, as will be shown later in the paper, the classical asymptotic theory does not apply to this class of estimators even for the mean-optimal criterion.

We propose a novel formulation of the estimator as a solution of an optimization problem with an estimated nuisance parameter. This representation allows us to further investigate the asymptotic theory of the estimated optimal treatment regime using empirical processes techniques. Our study reveals that the theory involves nonstandard asymptotics. We have rigorously established that: (1) the estimated parameter indexing the quantile-optimal treatment regime converges at a cube-root rate to a nonnormal limiting distribution that is characterized by the maximizer of a centered Gaussian process with a parabolic drift; and (2) the value function corresponding to the quantile optimal treatment regime can be estimated at an $O_p(n^{-1/2})$ rate. This new framework is broad in the sense that it also provides an alternative formulation of the mean optimal criterion, for which the same type of nonstandard asymptotics would arise. Thus, we fill an important gap in the literature. Moreover, the framework can be adapted to alternative criteria such as those based on weighted quantile or Gini's mean difference (Section 1.2 of online supplement). The main practical advantage of the proposed estimator is that it circumvents the difficulty of specifying a reliable outcome regression model, which has undue influence on estimating the optimal treatment regime.

We also investigate doubly robust estimation (Section 1.1 of online supplement), which can incorporate an outcome regression model when it is available.

In the causal inference context, several authors have considered estimating the quantile treatment effects for comparing several pre-determined treatment regimes (Rubin, 1974; Rosenbaum and Rubin, 1983; Hogan and Lee, 2004; Chernozhukov and Hansen, 2005; Zhang et al., 2012b). These authors have not investigated the fundamental problem of estimating the optimal treatment regimes in the quantile framework, which is much more complex than estimating the quantile specific treatment effect when the treatment assignment is given. Potentially, the recent work on discrete Q-learning in Moodie et al. (2014) can be applied to first estimate the probabilities and then invert them to estimate quantiles, but this application has not been systematically studied. Linn et al. (2015) independently considered estimating quantile-optimal treatment regime. However, their approach depends on applying threshold interactive model-based Q-learning at a sequence of thresholding values and then performing inversion. The method requires specifying the underlying outcome models and is computationally intensive even for homoscedastic error outcome models. Furthermore, Linn et al. (2015) has not studied the asymptotic theory we considered here.

The rest of the paper is organized as follows. The quantile-optimal treatment regime is proposed in Section 2. The estimation procedure and asymptotic distribution are introduced in Section 3. Section 4 investigates quantile-optimal dynamic treatment regimes. Simulation studies and a data example are reported in Section 5. Section 6 considers doubly robust estimation and alternative optimality criteria. The proofs are given in the Appendix. Additional technical details and numerical results can be found in the online supplement. The methods proposed in this paper can be implemented using the R package *quantoptr* (Zhou et al., 2017).

2 Quantile-optimal treatment regime

Let A be the binary variable denoting treatment (0 or 1 corresponding to two treatment options), and let Y denote the outcome. Without loss of generality, we assume that a larger value of the outcome is preferable. To evaluate the treatment effect, we consider the potential or counterfactual outcome framework (Neyman (1990), Rubin (1978)) for causal models. Let $Y^*(1)$ be the potential outcome had the subject been assigned to treatment 1; and $Y^*(0)$ be the potential outcome had the subject been assigned to treatment 0. For each individual in the sample, we observe either $Y^*(1)$ or $Y^*(0)$, but not both. It is assumed that the observed outcome is $Y = Y^*(1)A + Y^*(0)(1-A)$, that is, the observed outcome is the potential outcome corresponding to the treatment the subject actually receives. This is often referred to as the consistency assumption in causal inference. We also adopt the stable unit treatment value assumption (Rubin (1986)), that is, a subject's outcome of receiving a treatment is not influenced by the treatments received by other subjects.

Let X denote an l-dimensional vector of covariates. A treatment regime is defined as a function $d(X)$, that maps the covariates vector X to the set of treatment options, here $\{0, 1\}$. For example, $d(X) = I(X \leq 3/5)$ would assign a subject with $X = 0.2$ to treatment 1. Given treatment regime $d(X)$, the corresponding potential outcome is $Y^*(d) = Y^*(1)d(X) + Y^*(0)(1-d(X))$, that is, $Y^*(d)$ is the outcome one would observe if a subject with covariate value X is assigned to treatment 1 or 0 following treatment regime $d(X)$. We assume that $(Y^*(1), Y^*(0))$ is independent of A conditional on X (unconfoundedness assumption, Rosenbaum and Rubin (1983)), which is automatically satisfied in randomized trials.

Given a collection $\mathbb{D}$ of treatment regimes, the optimal treatment regime is typically defined as the one that maximizes the average of the potential outcome: $E(Y^*(d))$. Here, we consider a new quantile-optimal treatment regime, which is defined as

$$\arg max_{d \in \mathbb{D}} Q_\tau(Y^*(d)), \tag{1}$$

Table 1: Mean, 0.25 quantile and 0.10 quantile of the potential outcomes corresponding to six different treatment regimes (based on a Monte Carlo experiment with 10^6 observations).

Regime	(1)	(2)	(3)	(4)	(5)	(6)	(7)
mean	1.50	**2.40**	2.37	2.00	1.78	2.00	1.74
$Q_{0.25}$	0.80	1.10	**1.14**	1.01	0.91	-0.02	0.59
$Q_{0.10}$	0.16	-0.03	0.20	**0.33**	0.26	-2.29	-0.81

where $\tau \in (0,1)$ is the quantile level of interest and $Q_\tau(Y^*(d))$ is the τth quantile of $Y^*(d)$, specifically, $Q_\tau(Y^*(d)) = \inf\{t : F^*(t) \geq \tau\}$ with F^* denoting the distribution function of $Y^*(d)$.

To illustrate how the quantile-optimal treatment regime differs from the mean-optimal treatment regime, we consider a simple but instructive example. The outcome, Y_i, satisfies $Y_i = 1 + 3A_i + X_i - 5A_iX_i + (1 + A_i + 2A_iX_i)\epsilon_i$, where $\epsilon_i \sim N(0,1)$, $X_i \sim$ Uniform$[0,1]$, and $A_i = 1$ (or 0) if subject i receives treatment (or control). We consider the following six treatment regimes: (1) $A_i = 0, \ \forall \ i$; (2) $A_i = I(X_i \leq 3/5)$; (3) $A_i = I(X_i \leq 1/2)$; (4) $A_i = I(X_i \leq 1/5)$; (5) $A_i = I(X_i \leq 1/10)$; (6) $A_i = 1, \forall \ i$; and (7) random assignment $P(A_i = 1) = 0.5$. It is easy to derive that treatment regime 2 is the mean-optimal treatment regime. Table 1 summarizes the mean, the 0.25 quantile ($Q_{0.25}$) and 0.10 quantile ($Q_{0.10}$) of the potential outcome distribution corresponding to each of the six treatment regimes, based on a Monte Carlo experiment with 10^6 observations. We observe that regime 3 is the best if one is interested in maximizing the first quartile of the potential outcome distribution; whereas regime 4 performs best with respect to the 0.10 quantile. If we consider the hypothetical setting where the outcome is the survival time of cancer patients, then regime 2 (mean-optimal treatment regime) may have detrimental effect for patients at the lower tail, corresponding to weaker patients. Regime 3 significantly improves the survival time of the patients at the lower tail, while its mean value is comparable to that of regime 2. Thus, regime 3 is preferable if doctors wish to improve the life span of more severely ill patients without sacrificing the average treatment benefit of the population.

3 Estimation and large sample theory

3.1 Estimating quantile-optimal treatment regime

To explain the idea, we first consider a randomized trial with two treatment options (denoted by 1 and 0). Extensions to observational studies and dynamic treatment regimes will be discussed later. The observed data $\{X_i, Y_i, A_i\}$, $i = 1, \ldots, n$, are independent and identically distributed copies of $\{X, Y, A\}$. Our aim is to estimate the quantile-optimal treatment regime given a class of feasible treatment regimes $\mathbb{D} = \{I(X^T\beta{>}0) : \beta \in \mathbb{B}\}$, where β indexes different treatment regimes and $\mathbb{B}$ is a compact subset of $\mathbb{R}^l$. This class of *single-index* decision rules has been popular in practice (Zhang et al., 2012a, 2013; Zhao et al., 2012) due to its simplicity and interpretability. It is straightforward to show that this class contains the mean-optimal treatment regime corresponding to some popular choices of outcome models. For example, for the outcome model $E(Y|A, X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + A(\beta_3 + \beta_4 X_1 + \beta_5 X_2)$, the corresponding mean-optimal treatment regime is $I(\beta_3 + \beta_4 X_1 + \beta_5 X_2 > 0)$. An alternative class of treatment regimes that are practically appealing is the class of *thresholding rules* of the form $I(X_1 > \beta_1, \ldots, X_l > \beta_l)$, for some constants $\beta_1, \ldots, \beta_l$. Even for these relatively simple forms, asymptotic theory for the estimated optimal treatment regime, no matter what the criterion is, is nontrivial. It is worth pointing out that it is not necessary that the class of candidate treatment regimes includes the theoretically global optimal treatment regimes, as the interpretability of the treatment regime is often of fundamental importance.

We will focus on the single-index treatment regimes, as the theory for the thresholding decision rules is similar and simpler. Given a $\beta \in \mathbb{B}$, let $d(X, \beta) = I\{X^T\beta > 0\}$ be the treatment regime indexed by β, which is sometimes denoted by d_β for notational simplicity. For a quantile level of interest τ $(0 < \tau < 1)$, we would like to estimate $\beta_0 = \arg max_{\beta \in \mathbb{B}} Q_\tau(Y^*(d_\beta))$, the parameter indexing the quantile-optimal treatment regime. To do so, we make use of an induced missing data framework motivated by Zhang et al.

(2012a). Let $C(\beta) = Ad(X,\beta) + (1-A)(1-d(X,\beta))$. In the induced missing data framework, the "full data" of interest, but not completely observed, are $\{Y^*(d_\beta), X\}$; and the observed data are $\{C(\beta), C(\beta)Y^*(d_\beta), X\} = \{C(\beta), C(\beta)Y, X\}$. If $C(\beta) = 1$, then potential outcome $Y^*(d_\beta)$ is observed; if $C(\beta) = 0$ then $Y^*(d_\beta)$ is "missing". Furthermore, $Y^*(d_\beta)$ and $C(\beta)$ are independent conditional on X. Thus, the induced missing data structure satisfies the missing at random assumption. Let

$$\widehat{Q}_\tau(\beta) = \arg\min_a n^{-1} \sum_{i=1}^{n} C_i(\beta)\rho_\tau(Y_i - a), \tag{2}$$

where $\rho_\tau(u) = u(\tau - I(u<0))$ is the quantile loss function. As stated in the following lemma (proof given in the online supplement), $\widehat{Q}_\tau(\beta)$ is a consistent estimator of the τth quantile of $Y^*(d_\beta)$.

Lemma 1. *If condition (C1) in Section 3.3 is satisfied, then we have* $\widehat{Q}_\tau(\beta) \to Q_\tau(Y^*(d_\beta))$ *in probability,* $\forall\ \beta \in \mathbb{B}$.

The estimator for β_0 that corresponds to the quantile-optimal treatment regime is

$$\widehat{\beta}_n = \operatorname*{argmax}_{\beta\in\mathbb{B}} \widehat{Q}_\tau(\beta). \tag{3}$$

The estimated quantile-optimal treatment regime is $d_{\hat{\beta}} = I(X^T\widehat{\beta} > 0)$. Section 2.1 of the online supplement provides the calculation details.

3.2 Alternative formulation of the proposed estimator

As the treatment regimes involve indicator functions, the nonsmoothness leads to nonstandard asymptotics even when the mean criterion is used. The asymptotic theory is challenging and involves a cube-root convergence rate and a non-normal limiting distribution, see Section 3.3 for details. Even for the mean criterion, the asymptotic distribution theory of

the estimated optimal treatment regime has not yet been systematically developed in the literature.

To facilitate the development of theory, we introduce a novel reformulation that represents the quantile-optimal treatment regime parameter estimator (3) as a solution of an optimization problem with an estimated nuisance parameter. To motivate the reformulation, let

$$
\begin{aligned}
g(\cdot, \beta, m) &= C(\beta) I\{Y - m > 0\}, && (4)\\
m_0 &= \sup\{m : \sup_{\beta \in \mathbb{B}} Pg(\cdot, \beta, m) \geq (1-\tau)/2\}, && (5)\\
\beta_0 &= \operatorname*{argmax}_{\beta \in \mathbb{B}} Pg(\cdot, \beta, m_0). && (6)
\end{aligned}
$$

The function $g(\cdot, \beta, m)$ is motivated by the first-order condition of the maximization problem in (2). For a randomized trial, $P(C(\beta) = 1|X) = P(C(\beta) = 0|X) = \frac{1}{2}$. Thus, $P(g(\cdot, \beta, m)) = \frac{1}{2} P(Y^*(d_\beta) > m)$, which is equal to $\frac{1-\tau}{2}$ if $m = Q_\tau(Y^*(d_\beta))$. For any given β, because $g(\cdot, \beta, m)$ is monotonically decreasing in m, it follows that $Q_\tau(Y^*(d_\beta))$ is the largest value of m such that $Pg(\cdot, \beta, m)$ is greater than or equal to $\frac{1-\tau}{2}$. Therefore, m_0 defined in (5) is the largest achievable τth quantile of $Y^*(d_\beta)$ over $\beta \in \mathbb{B}$; and β_0 defined in (6) is the population value of the parameter that indexes the optimal treatment regime.

Now, let P_n denote the empirical expectation, that is, $P_n f(Z) = n^{-1} \sum_{i=1}^n f(Z_i)$, where $Z_1, \ldots, Z_n$ is a random sample and $f(\cdot)$ is an arbitrary function. Then, $\widehat{m}_n = \sup\{m : \sup_{\beta \in \mathbb{B}} P_n g(\cdot, \beta, m) \geq (1-\tau)/2\}$ is the estimator of the largest achievable τth quantile of $Y^*(d_\beta)$ over the class of treatment regimes under consideration. We have the following alternative expression of the estimator in (3):

$$
\widehat{\beta}_n = \operatorname*{argmax}_{\beta \in \mathbb{B}} P_n g(\cdot, \beta, \widehat{m}_n). \qquad (7)
$$

In other words, $\widehat{\beta}_n$ is the value of β at which the supremum of $P_n g(\cdot, \beta, \widehat{m}_n)$ is achieved,

thus it is the estimator of the parameter that indexes the optimal treatment regime. This reformulation was partly motivated by the least median of squares estimator of Rousseeuw (1984). A benefit of this reformulation is that we also obtain the convergence rate of $\widehat{m}_n$, which is the estimator for the maximally achievable value function (here, the maximally achievable τth quantile of the potential outcome) as a by product (see Lemma 2 in Section 3.3).

3.3 Asymptotic properties

We assume the following regularity conditions.

(C1) Potential outcomes $Y^*(1)$ and $Y^*(0)$ both have continuous distributions with bounded, continuously differentiable density functions.

(C2) The population parameter indexing the optimal treatment regime, $\beta_0 \in \mathbb{R}^l$, which satisfies $||\beta_0|| = 1$, where $||\cdot||$ denotes the Euclidean norm, is unique and is an interior point of $\mathbb{B}$, a compact subset of the parameter space.

(C3) X has a continuously differentiable density function $f(\cdot)$. The angular components of X, considered as a random element of the unit sphere $\mathbb{S}$ in $\mathbb{R}^l$, has a bounded, continuous density with respect to the surface measure on $\mathbb{S}$.

(C4) Let $q(X, \delta) = S_{1,X}(m_0 + \delta) - S_{0,X}(m_0 + \delta)$, where $S_{1,X}(\cdot)$ and $S_{0,X}(\cdot)$ denote the conditional survival functions of $Y^*(1)$ and $Y^*(0)$ given X, respectively; and $\dot{q}(X, 0)$ and $\dot{f}(X)$ denote the gradients with respect to X. The $l \times l$ matrix $V = \frac{1}{2} \int I\{X^T\beta_0 = 0\}(f(X)\dot{q}(X, 0) + q(X, 0)\dot{f}(X))'\beta_0 X X^T d\sigma$ is positive definite, where σ is the surface measure on the hyperplane $\{X : X^T\beta_0 = 0\}$.

Condition (C1) is a standard assumption on the potential outcomes in causal inference. Condition (C2) is an identifiability condition for β_0. Conditions (C3) and (C4) are technical conditions to evaluate the quadratic drift and covariance function of the Gaussian process

that are used to characterize the asymptotic distribution of $\widehat{\beta}_n$. The matrix V in (C4) characterizes the quadratic drift of the Gaussian process. These two conditions are similar to those in Example 6.4 of Kim and Pollard (1990). In particular, condition (C3) is mainly imposed for the convenience of calculating the derivative of the surface integral in the proof of Lemma 2. It can be relaxed to allow some of the components of X to be discrete at the expense of a more complex expression for V. The new formulation in Section 3.2 connects the problem of estimating β_0 to the class of estimation problems with cube root asymptotics (Kim and Pollard, 1990). However, the result of Kim and Pollard (1990) is not directly applicable because our estimator of β_0 contains an estimated nuisance parameter $\widehat{m}_n$. Lemma 2 below shows that $\widehat{\beta}_n$ nearly maximizes the objective function in (7) in which $\widehat{m}_n$ is replaced by the limiting value m_0.

Lemma 2. *Under conditions (C1)-(C4),*

(1) $\widehat{m}_n = m_0 + O_p(n^{-1/2})$.

(2) $P_n g(\cdot, \widehat{\beta}_n, m_0) \geq \sup_{\beta \in \mathbb{B}} P_n g(\cdot, \beta, m_0) - o_p(n^{-2/3})$.

The first part of Lemma 2 shows that $\widehat{m}_n$ has a root-n convergence rate. This result is of independent interest as it tells us how well we could estimate the theoretically largest achievable value of the criterion function. The details of the derivation of the lemma are given in the Appendix. Lemma 2 confirms that $\widehat{\beta}_n$ nearly maximizes $P_n g(\cdot, \beta, m_0)$. This allows us to further derive the asymptotic distribution of $\widehat{\beta}_n$, which is expressible as a functional of two-sided Brownian motion with a quadratic drift. This result is stated in the following theorem.

Theorem 1. *Assume conditions (C1)-(C4) are satisfied. Then,* $n^{1/3}(\widehat{\beta}_n - \beta_0)$ *converges in distribution to* $\arg max_t Z(t)$*, where the process* $Z(t) = -\frac{1}{2}t^T V t + W(t)$*,* V *is an* $l \times l$ *positive definite matrix and* $W(t)$ *is a centered Gaussian process with continuous sample paths and covariance kernel function* $K(\cdot, \cdot)$*. The expressions for* V *and* $K(\cdot, \cdot)$ *are given in the proof of the theorem in the Appendix.*

Remark 1. If the mean-optimal criterion is of interest, then we let $g^*(\cdot, \beta, \mu) = C(\beta)(Y - \mu)$ and $\widehat{\mu}_n = \sup\{\mu : \sup_{\beta \in \mathbb{B}} P_n g^*(\cdot, \beta, \mu) > 0\}$. The estimated parameter indexing the mean-optimal treatment regime has the representation $\widehat{\beta}_n^{\text{mean}} = \underset{\beta \in \mathbb{B}}{\text{argmax}}\ P_n g^*(\cdot, \beta, \widehat{\mu}_n)$. It is straightforward to adapt the techniques developed in this paper to show that the estimated parameter indexing the mean-optimal treatment regime has a non-standard convergence rate and a non-normal limiting distribution. This fills an important gap in the literature.

Remark 2. If the observed data arise from observational studies, the above formulation and theory can be extended using propensity score weighting. For observational studies, we have $Y^*(d_\beta) \perp C(\beta) \big| X$, which is guaranteed under the common causal inference assumption $\{Y^*(1), Y^*(0)\} \perp A \big| X$. Thus, the "missing at random" assumption is satisfied in the induced missing data framework of Section 3.1. Let $\pi(X) = P(A = 1|X)$, then the propensity score $P(C_\beta = 1 \big| X)$ has the expression $\pi(X)d(X, \beta) + (1 - \pi(X))(1 - d(X, \beta))$. We denote the propensity score by $\pi_c(X, \beta)$ for notational simplicity. We then estimate the τth quantile of $Y^*(d_\beta)$ by $\tilde{Q}_\tau(\beta) = \underset{a}{\text{argmin}} n^{-1} \sum_{i=1}^n \frac{C_i(\beta)}{\hat{\pi}_c(X_i, \beta)} \rho_\tau(Y_i - a)$, where $\hat{\pi}_c(X_i, \beta)$ is an estimator of the propensity score $\pi_c(X, \beta)$. A simple way to obtain $\hat{\pi}_c(X_i, \beta)$ is to estimate $\pi(X)$ based on $\{A_i, X_i\}$, $i = 1, \ldots, n$, using logistic regression, which models $\pi(X)$ as $\pi(X, \gamma) = \exp(X^T\gamma)/(1 + \exp(X^T\gamma))$. One may also use semiparametric or nonparametric models, which renders greater flexibility but demands heavier computation. The estimated parameter indexing the quantile-optimal treatment regime is given by $\underset{\beta \in \mathbb{B}}{\text{argmax}}\ \tilde{Q}_\tau(\beta)$.

4 Quantile-optimal dynamic treatment regimes

When treating chronic medical conditions, it is frequently necessary to vary the treatment (e.g., drug type, dose) over time according to how the patient responds to the previous treatment. This motivates us to consider estimating the quantile-optimal dynamic treatment regime (DTR) using data from longitudinal studies, which can also help catch possible delayed treatment effects. Comparing with the static treatment regime discussed earlier,

a new challenge is the presence of time-dependent covariates that may be simultaneously confounders and intermediate variables.

Consider a two-stage longitudinal study in which the subject receives treatment $A_1 \in \{0,1\}$ at stage 1 and treatment $A_2 \in \{0,1\}$ at stage 2. We are interested in the outcome at the end of the study. We would like to estimate the optimal DTR $d = (d_1, d_2)$, where d_j can depend on the realized covariates and treatment history before the jth decision, $j = 1, 2$. The baseline vector of covariates is denoted by X_1, the potential outcomes are denoted by $\{X_2^*(d_1), Y^*(d)\}$, where $X_2^*(d_1)$ is the covariate information between decisions d_1 and d_2 had the subject received treatment d_1, and $Y^*(d)$ is the potential outcome had the subject received treatment $d = (d_1, d_2)$. As before, we define the quantile-optimal DTR as $d^{\text{opt}} = \underset{d \in \mathbb{D}}{\text{argmax}} Q_\tau(Y^*(d))$. Let $H_1 = \{X_1\}$ and $H_2 = \{X_1, A_1, X_2\}$. We adopt the no unmeasured confounder or sequential ignorability assumption (Robins (1997)), that is, given any regime (a_1, a_2), $A_1 \perp \{X_2^*(a_1), Y^*(a_1, a_2)\} | H_1$ and $A_2 \perp Y^*(a_1, a_2) | H_2$. In other words, treatment A_j received in the jth stage ($j = 1, 2$) is independent of any future (potential) covariate or outcome conditional on the history. We also adopt the positivity assumption, that is, there exist positive constants $c_1 < c_2$ such that $c_1 \leq P(A_j = a | H_j) \leq c_2$, with probability one, for $a \in \{0,1\}$, $j = 1, 2$. Assume that the class of candidate treatment regimes is indexed by $\xi = (\beta^T, \gamma^T)^T \in \mathbb{B} = \mathbb{B}_1 \times \mathbb{B}_2$, $d_\xi = (d_\beta, d_\gamma)$, where $d_\beta(H_1) = I(H_1^T \beta > 0)$ and $d_\gamma(H_2) = I(H_2^T \gamma > 0)$.

The observed data are denoted by $\{X_{i1}, A_{i1}, X_{i2}, A_{i2}, Y_i\}$, $i = 1, \ldots, n$, where X_{i1} denotes the baseline vector of covariates for subject i, A_{i1} is the treatment subject i receives at stage 1, X_{i2} denotes the vector of intermediate information observed between the two stages, A_{i2} is the treatment subject i receives at stage 2, and Y_i is the observed outcome for subject i (as before, a larger value is preferred). To estimate the optimal treatment regime, we consider a similar induced missing data structure, as motivated by Zhang et al. (2013). For a given treatment regime d_ξ, the "full data" are $(X_1, X_2^*(d_\beta), Y^*(d_\xi))$. Let $C_\xi = \infty$ if $A_1 = d_\beta$ and

$A_2 = d_\gamma$. In this case, $(X_1, X_2, Y) = (X_1, X_2^*(d_\beta), Y^*(d_\xi))$, and we observe the potential outcome. Let $C_\beta = 2$ if $A_1 = d_\beta$ but $A_2 \neq d_\xi$ (dropout before decision 2); and let $C_\beta = 1$ if $A_1 \neq d_\beta$ and $A_2 \neq d_\xi$ (dropout before decision 1). Note that this induced missing data structure mimics the monotone dropout scenario for longitudinal data. We can verify that the setup satisfies the missing at random assumption, that is, missingness may be related to the observed information but is conditionally independent of what is missing.

Let $\pi_1(H_1) = P(A_1 = 1 \mid H_1)$ and $\pi_2(H_2) = \pi_2(\overline{X}_2, a_2) = P(A_2 = 1 \mid \overline{X}_2, a_2)$, where $\overline{X}_2 = (X_1^T, X_2^T)^T$ is an l-dimensional vector. It is important to note that H_2 depends on the treatment received at the first stage. If the subject receives $A_1 = a_1 \in \{0, 1\}$ at the first stage, we sometimes write H_2 as $H_2(a_1) = \{X_1, a_1, X_2\}$ to emphasize the dependence, for which case $X_2 = X_2^*(a_1)$ by the consistency assumption. Similarly, for $A_1 = d_\beta(H_1)$, we sometimes write H_2 as $H_2(d_\beta) = \{d_\beta(X_1), X_2\}$. The potential outcomes correspond to d_ξ are denoted by $\{X_1, X_2^*(d_\beta(X_1)), Y^*(d_\xi)\}$ or simply $\{X_2^*(d_\beta), Y^*(d_\xi)\}$.

As before, we would minimize $P_n\Big(\frac{I(C_\xi=\infty)}{P(C_\xi=\infty\big|H_2)}\rho_\tau(Y-a)\Big)$ to estimate the τth quantile of $Y^*(d_\xi)$. Note that $C_\xi = \infty$ if and only if $A_1 = d_\beta(X_1)$ and $A_2 = d_\gamma(H_2(d_\beta))$, in other words, $H_2 = H_2(d_\beta)$ or the observed history is the potential history corresponding to d_β. Thus, in the above inverse probability weighted quantile loss function

$$\begin{aligned} P(C_\xi = \infty\big|H_2) &= P(C_\xi = \infty\big|X_1, X_2^*(d_\beta(X_1))) \\ &= P(A_1 = d_\beta\big|X_1, X_2^*(d_\beta(X_1)))P(A_2 = d_\gamma\big|A_1 = d_\beta(X_1), X_1, X_2^*(d_\beta(X_1))) \\ &= P(A_1 = d_\beta\big|X_1)P(A_2 = d_\gamma\big|H_2(d_\beta)) \end{aligned}$$

where $P(A_1 = d_\beta\big|X_1) = \big[\pi_1(H_1)d_\beta + (1-\pi_1(H_1))(1-d_\beta)\big]$ and $P(A_2 = d_\gamma\big|H_2(d_\beta)) = \big[\pi_2(H_2(d_\beta))d_\gamma + (1-\pi_2(H_2(d_\beta)))(1-d_\gamma)\big]$. For notational simplicity, we denote $P(C_\xi = \infty\big|H_2)$ by $\pi(\xi)$. Formally, the estimate of the τth quantile of $Y^*(d_\xi)$ is given by $\widehat{Q}_\tau(\xi) = \underset{a}{\text{argmin}} n^{-1}\sum_{i=1}^n \frac{I(C_{\xi,i}=\infty)}{\pi(\xi)}\rho_\tau(Y_i - a)$, where $C_{\xi,i}$ is the value of C_ξ for subject i. The consistency of $\widehat{Q}_\tau(\xi)$ is established in the online supplement. The estimator of the parameter indexing the

optimal DTR from the class $\mathbb{D}$ is defined as $\widehat{\xi} = \underset{\xi=(\beta^T,\gamma^T)^T\in\mathbb{B}}{\text{argmax}}\ \widehat{Q}_\tau(\xi)$. The estimated quantile-optimal treatment regime is $d_{\hat{\xi}} = (d_{\widehat{\beta}}, d_{\widehat{\gamma}})$.

In the following, we assume that the data arise from a SMART (sequential, multiple, assignment randomized trials), which has been recommended as a standard design for optimal DTR estimation (Lavori and Dawson, 2000; Murphy, 2008). For a SMART, $\pi_1(H_1)$ and $\pi_1(H_2)$ are both known by design, thus $\pi(\xi)$ is known for any given ξ. Let $g(\cdot,\xi,m) = \frac{I(C_\xi=\infty)}{\pi(\xi)} I(Y > m)$ and $\widehat{m}_n = \sup\{m : \sup_\xi P_n g(\cdot,\xi,m) \geq (1-\tau)\}$. We have the following alternative expression: $\widehat{\xi}_n = \underset{\xi}{\text{argmax}}\ P_n g(\cdot,\xi,\widehat{m}_n)$. Let $m_0 = \sup\{m : \sup_\xi Pg(\cdot,\xi,m) \geq (1-\tau)\}$ and $\xi_0 = \underset{\xi}{\text{argmax}}\ Pg(\cdot,\xi,m_0)$. Under similar conditions as for Theorem 1, it can be derived that the limiting distribution of $n^{1/3}(\widehat{\xi}_n - \xi_0)$ is that of the maximizer of a centered Gaussian process with a quadratic drift.

Theorem 2. *Under conditions (C1*)–(C4*) given in the online supplement,* $n^{1/3}(\widehat{\xi}_n - \xi_0)$ *converges in distribution to* $\arg max_t Z^*(t)$*, where the process* $Z^*(t) = -\frac{1}{2}t^T V^* t + W^*(t)$*,* V^* *is an* $l \times l$ *positive definite matrix and* $W^*(t)$ *is a centered Gaussian process with continuous sample paths and covariance kernel function* $K^*(C_1, C_2)$*. The expressions for* V^* *and* $K^*(\cdot,\cdot)$ *are given in the online supplement.*

5 Numerical results

5.1 Simulations

Example 1 (single-stage optimal treatment regime). We compare estimating the conventional mean-optimal treatment regime and quantile-optimal treatment regime in this example. We generate random data from the model $Y = 1 + X_1 - X_2 + X_3^3 + e^{X_4} + A(3 - 5X_1 + 2X_2 - 3X_3 + X_4) + (1 + A(1 + X_1 + X_2 + X_3 + X_4))\epsilon$, where X_k $(k = 1, ..., 4)$ are independent Uniform$(0, 1)$ random variables and $\epsilon \sim N(0, 1)$ is independent of the covariates. The binary treatment indicator A satisfies $\log\left(P\left(A = 1\middle|X\right)/P\left(A_i = 0\middle|X\right)\right) =$

Table 2: Population parameters and summary values for optimal treatment regimes under different criteria for Example 1 based on a Monte Carlo experiment with $n = 10^5$.

	η_0	η_1	η_2	η_3	η_4	Qmean	$Q_{0.25}$	$Q_{0.1}$
mean criterion	0.43	-0.72	0.29	-0.43	0.14	3.99	2.28	0.55
0.25qt criterion	0.42	-0.60	0.41	-0.43	-0.34	3.79	2.46	1.18
0.1qt criterion	0.27	-0.68	0.38	-0.43	-0.37	3.44	2.36	1.55

Columns 2-6 are values of the η_i's of the optimal treatment regimes corresponding to different criteria. The last three columns are the mean, 0.25 quantile and 0.1 quantile of the potential outcomes if the optimal treatment regime is applied.

$-0.5 - 0.5\,(X_1 + X_2 + X_3 + X_4)$, where $X = (X_1, \ldots, X_4)'$.

We consider the class of treatment regimes $I(\eta_0 + \eta^T X > 0)$, where $(\eta_0, \eta_1, \ldots, \eta_4)^T$ has L_2-norm 1. Let $\mu(a, X) = E(Y|A = a, X)$, where $a \in \{0, 1\}$. The mean optimal treatment regime is given by $I(\mu(1, X) > \mu(0, X))$. In our example, it is $I(3 - 5X_1 + 2X_2 - 3X_3 + X_4 > 0)$, which belongs to our class of candidate treatment regimes. We compare the proposed method with two popular methods for estimating the mean-optimal treatment regime: a model-based approach and a model-free approach. For the model-based approach we impose models for $\mu(a, X)$ and then estimate the mean-optimal treatment regime by $I(\hat{\mu}(1, X) > \hat{\mu}(0, X))$, where $\hat{\mu}$ is the estimated value of μ. We consider two posited models for $\mu(a, X)$: (1) correctly specified regression function $\mu_t\,(a, X) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3^3 + \gamma_4 e^{X_4} + a\,(\gamma_5 + \gamma_6 X_1 + \gamma_7 X_2 + \gamma_8 X_3 + \gamma_9 X_4)$; and (2) misspecified regression function $\mu_m\,(a, X) = \exp\left[\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3^3 + a\,(\gamma_4 + \gamma_5 X_1 + \gamma_6 X_2 + \gamma_7 X_3 + \gamma_8 X_4)\right]$. For the model-free approach, we consider the estimator in Zhang et al. (2012a). We denote these three estimators by mean$_RG_{\mu_t}$, mean$_RG_{\mu_m}$ and mean$_$ZTLD, respectively.

For the quantile criteria, we consider $\tau = 0.25$ and 0.1, and denote the corresponding criterion as 0.25qt criterion and 0.10qt criterion, respectively. We do not have a closed form expression for the quantile-optimal treatment regime. In Table 2, based on a Monte Carlo experiment with sample size 10^5, we report the values of the η_i's indexing the optimal treatment regimes corresponding to different criteria; the last three columns of the table

Table 3: Estimated optimal treatment regimes (mean criterion, 0.25 quantile criterion and 0.1 quantile criterion) and their corresponding value functions for Example 1.

Method	n	$\hat{\eta}_0$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\eta}_4$	$\hat{Q}_{\text{mean}}$	$\hat{Q}_{0.25}$	$\hat{Q}_{0.1}$
mean_RG_{μ_t}	500	0.42	-0.71	0.28	-0.41	0.14	3.99	2.29	0.56
		(0.10)	(0.07)	(0.13)	(0.11)	(0.12)	(0.21)	(0.19)	(0.40)
	1000	0.43	-0.71	0.29	-0.43	0.14	3.99	2.28	0.52
		(0.06)	(0.05)	(0.09)	(0.08)	(0.09)	(0.14)	(0.13)	(0.27)
mean_RG_{μ_m}	500	0.26	-0.71	0.30	-0.38	0.37	3.96	2.23	0.65
		(0.11)	(0.08)	(0.12)	(0.12)	(0.12)	(0.21)	(0.19)	(0.38)
	1000	0.27	-0.71	0.31	-0.39	0.37	3.97	2.22	0.62
		(0.08)	(0.06)	(0.09)	(0.08)	(0.09)	(0.15)	(0.13)	(0.27)
mean_ZTLD	500	0.36	-0.63	0.31	-0.38	0.12	4.31	2.31	0.63
		(0.2)	(0.14)	(0.24)	(0.2)	(0.27)	(0.21)	(0.21)	(0.53)
	1000	0.38	-0.67	0.29	-0.4	0.17	4.18	2.29	0.6
		(0.15)	(0.11)	(0.19)	(0.15)	(0.19)	(0.13)	(0.16)	(0.47)
0.25qt criterion	500	0.38	-0.57	0.37	-0.37	-0.31	3.85	2.65	1.3
		(0.15)	(0.14)	(0.19)	(0.18)	(0.2)	(0.26)	(0.16)	(0.39)
	1000	0.4	-0.59	0.35	-0.43	-0.28	3.81	2.57	1.31
		(0.12)	(0.12)	(0.17)	(0.12)	(0.15)	(0.18)	(0.11)	(0.28)
0.10qt criterion	500	0.24	-0.56	0.3	-0.4	-0.33	3.5	2.45	1.75
		(0.23)	(0.2)	(0.25)	(0.22)	(0.25)	(0.26)	(0.16)	(0.15)
	1000	0.27	-0.61	0.32	-0.44	-0.33	3.47	2.42	1.68
		(0.18)	(0.14)	(0.22)	(0.15)	(0.19)	(0.18)	(0.11)	(0.11)

The numbers in the parenthesis are standard deviations. The last three columns are the estimated mean, 0.25 quantile and 0.1 quantile of the potential outcome if the estimated optimal treatment regime is applied. The three methods mean_RG_{μ_t}, mean_RG_{μ_m} and mean_ZTLD denote the mean-optimal treatment regime estimators using the model-based approach with correctly specified regression model, the model-based approach with incorrectly specified regression model and the approach of Zhang et al. (2012a), respectively.

contain the mean, the 0.25 quantile and the 0.1 quantile of the outcomes if the corresponding optimal treatment regime is applied. These values will serve as our gold standard.

Table 3 summarizes the estimated optimal treatment regimes corresponding to the mean criterion (using mean_RG_{μ_t}, mean_RG_{μ_m} and mean_ZTLD, respectively), the 0.25qt criterion and the 0.10qt criterion for sample sizes n =500 and 1000. The last three columns of Table 3 report the estimated mean, the 0.25 quantile and the 0.1 quantile of the outcomes if the estimated optimal treatment regime is applied. We observe the model-based approach for estimating the mean-optimal treatment regime is sensitive to the specified regression model and can be biased when the regression model is misspecified (mean_RG_{μ_m} gives very biased estimators for η_0 and η_4). Also, the estimated optimal treatment regimes (and their achievable performance in terms of the value of the criterion functions) using the model-free approach get closer to the ideal ones reported in Table 2 as the sample size increases.

Example 2 (two-stage DTR). We generate random data from the following model $Y = 1+X_1+A_1\left[1-3\left(X_1-0.2\right)^2\right]+X_2+A_2\left[1-5\left(X_2-0.4\right)^2\right]+(1+0.5A_1-A_1X_1+0.5A_2-A_2X_2)\epsilon$, where $\epsilon \sim N(0, 0.4)$, $X_1 \sim \text{Uniform}(0,1)$, $X_2|\{X_1, A_1\} \sim \text{Uniform}(0.5X_1, 0.5X_1+0.5)$, $A_1|X_1 \sim \text{Bernoulli}(\text{expit}(-0.5+X_1))$, and $A_2|\{X_1, A_1, X_2\} \sim \text{Bernoulli}(\text{expit}(-1+X_2))$ with $\text{expit}(t) = e^t/(1+e^t)$. We consider sequential treatment regimes of the form (A_1, A_2), where $A_1 = I\{X_1 < \eta_1\}$, and $A_2 = I\{X_2 < \eta_2\}$. We note that this class contains the mean-optimal sequential treatment regimes which are given by $A_1 = I(X_1 < 0.777)$ and $A_2 = I(X_2 < 0.847)$.

The popular Q-learning procedure relies on specification of models for the so-called Q-functions. In this example, we compare with standard application of Q-learning based on linear models, that is, the Q-functions are specifies as $Q_t(H_t, A_t; \beta_t) = H_{t,0}^T\beta_{t,0} + A_tH_{t,1}^T\beta_{t,1}$, $t = 1, 2$, where $H_{2,0} = (1, X_1, A_1, X_1A_1, X_2)^T$, $H_{2,1} = (1, X_2)^T$, $H_{1,0} = (1, X_1)^T$, and $H_{1,1} =$

Table 4: Population parameters and summary values for optimal treatment regimes under different criteria for Example 2 based on a Monte Carlo experiment with $n = 10^5$.

Method	η_1	η_2	Q_{mean}	$Q_{0.50}$	$Q_{0.75}$
Mean criterion	0.777	0.847	3.331	3.323	3.821
0.50qt criterion	0.753	0.808	3.327	3.327	3.827
0.75qt criterion	0.729	0.795	3.322	3.325	3.828

Columns 2-3 are values of the η_i's of the optimal treatment regimes corresponding to different criteria. The last three columns are the mean, median and 0.75 quantile of the potential outcomes if the optimal treatment regime is applied.

$(1, X_1)^T$. We note that in practice the Q-learning procedure usually misspecifies the Q-function. We also compare with the model-free approach for estimating the mean-optimal dynamic treatment regime (Zhang et al. (2013)).

Table 4 reports the parameters indexing the optimal treatment regimes and the corresponding mean, median and 0.75 quantile of the outcome if the optimal treatment regime is applied, based on a Monte Carlo experiment with sample size 10^5. Table 5 summarizes the estimated parameters indexing the optimal treatment regimes and their estimated performance corresponding to different criteria for sample sizes $n = 500, 1000$, based on 400 simulation runs. The estimated optimal treatment regimes and their achievable performance are quite close to the ideal ones reported in Table 4, particularly when the sample size is large.

5.2 ACTG175 data analysis

We illustrate the proposed quantile-optimal treatment regime estimation method on the ACTG175 data set from the R package `speff2trial`, which contains measurements on 2139 HIV-infected patients. The patients were randomized to four treatment arms: zidovudine (AZT) monotherapy, AZT+didanosine (ddI), AZT+zalcitabine(ddC), and ddI monotherapy. The goal of the original clinical trial was to evaluate whether treatment of HIV infection with one drug (monotherapy) was the same as, better than, or worse than treatment with two drugs (combination therapy) in patients with CD4 T cells between 200 and $500/mm^3$

Table 5: Estimated optimal treatment regimes and their corresponding estimated value functions under different criteria for Example 2.

Method	n	η_1	η_2	$\hat{Q}_{mean}$	$\hat{Q}_{0.50}$	$\hat{Q}_{0.75}$
mean_Qlearning	500	0.755(0.041)	1.176(0.294)	3.319(0.090)	3.309(0.102)	3.815(0.122)
	1000	0.752(0.027)	1.131(0.144)	3.321(0.065)	3.305(0.07)	3.819(0.079)
mean_ZTLD	500	0.773(0.073)	0.846(0.067)	3.370(0.095)	3.376(0.097)	3.862(0.118)
	1000	0.768(0.055)	0.852(0.059)	3.356(0.065)	3.354(0.068)	3.848(0.081)
0.50qt criterion	500	0.751(0.08)	0.815(0.079)	3.357(0.090)	3.391(0.102)	3.858(0.119)
	1000	0.750(0.062)	0.813(0.069)	3.343(0.063)	3.366(0.068)	3.849(0.081)
0.75qt criterion	500	0.734(0.108)	0.785(0.103)	3.328(0.095)	3.331(0.109)	3.892(0.123)
	1000	0.723(0.084)	0.795(0.095)	3.322(0.067)	3.326(0.075)	3.865(0.077)

The numbers in the parenthesis are standard deviations. The last three columns are the estimated mean, median and 0.75 quantile of the potential outcome if the estimated optimal treatment regime is applied. The mean_Qlearning method stands for the mean-optimal treatment regime estimator using the Q-learning approach. The mean_ZTLD method is the mean-optimal treatment regime estimator using Zhang et al. (2013).

(Hammer et al., 1996). Figures 1 and 2 of the online supplement display the histograms of the response variable (CD4 count at week 96) for each of the two treatment arms for different subgroups of patients for which the subgroups are formed by the observed values of the CD4 count at week 0 or baseline weight. The varying shapes of the histograms across different ranges of both covariates indicate heteroscedastic treatment effects. It is also observed that the distribution of the response variable tends to be asymmetric and skewed to the right.

A basic conclusion from the study is for patients who had taken AZT before entering the trial, treatments with ddI or AZT + ddI are better than continuing to take AZT alone. There are $n = 562$ patients with full CD4 information that had taken AZT before the study and received AZT+ddI or ddI monotheraphy in this trial. Motivated by the aforementioned finding, we consider the problem of how to assign treatment to the patients who had taken AZT before, either to the AZT+ddI combination therapy or to the ddI monotheraphy. The quantitative outcome is the CD4 count at 96± 5 weeks from baseline (denoted as cd496) as CD4 count represents a vital signal for disease progression for HIV-infected patients. The treatment indicator A_i is set to 1 if patient i is assigned to the AZT + ddI therapy, and A_i is set to 0 if the patient is assigned to the ddI monotheraphy. Because this trial is randomized,

Table 6: Estimated optimal treatment regimes and summary values for ACTG175 data analysis.

Method	$\hat{\eta}_0$	$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{Q}_{0.50}$	$\hat{Q}_{0.25}$	$\hat{Q}$mean
0.50qt criterion	-0.571	0.691	0.444	360	220	375.4
0.25qt criterion	-0.210	0.958	-0.194	333	263	346.5
Mean criterion	-0.526	0.799	0.292	331	219	403.9

the propensity score $\pi_i = n^{-1}\sum A_i = 0.48$ is taken as a constant for all subjects.

Two covariates are considered for estimating the optimal treatment regimes: X_1 (baseline weight of patient, measured in kg) and X_2 (baseline CD4 T cell count, denoted by cd40). It has been observed that body weight has a significant role on AZT pharmacokinetic profile. Burger et al. (1994) reported that AZT clearance is significantly lower in patients with a lower body weight, which indicates a qualitative interaction with AZT. In medicine, drug clearance is a pharmacokinetic measurement of the rate at which the active drug is removed from the body, and drug clearance is correlated with the time course of a drug's action (Hammer et al., 1996).

Let $X = (X_1, X_2)$, where both X_1 and X_2 are standardized to the interval [0,1]. We consider the class of candidate regimes of the form $I\{\eta_0 + \eta_1 X_1 + \eta_2 X_2 < 0\}$, where $(\eta_0, \eta_1, \eta_2) \in (-1, 1)^3$. When the decision rule takes the value 1, the patient is assigned to the AZT+ddI combination therapy; otherwise the patient is assigned to the ddI monotheraphy. For identifiability, we impose the restriction $\|\eta\| = 1$. We estimate the optimal treatment regimes using the median criterion, quartile criterion and the mean criterion. The median criterion is motivated by the robustness consideration; the quartile criterion is motivated by the desire to improve the treatment effect for weaker patients. Table 6 summaries the estimated optimal treatment regimes for the three criteria.

The estimated median of the potential outcome when the median-optimal treatment regime is applied is 360; whereas the median of the observed outcome is 339.5. The estimated first quartile of the potential outcome when the 0.25qt criterion is applied is 263; whereas

the 0.25 quartile of the observed outcome is 237. The estimated mean of the potential outcome when the mean-optimal treatment regime is applied is 403.9; whereas the mean of the observed outcome is 355. Figure 3 of the online supplement depicts the three estimated regimes graphically, from which we observe that they are dramatically different from each other.

6 Conclusions and discussions

In a variety of applications, it is of interest to consider a treatment regime that maximizes the median or other quantile of the potential outcome distribution. This paper studies robust estimation of quantile-optimal static/dynamic treatment regimes. We propose a novel representation that expresses the parameter indexing the optimal treatment regime as a solution to an optimization problem with a nuisance parameter. Employing this representation and empirical process techniques, we prove that the estimated parameter indexing the quantile-optimal treatment regime has a nonstandard convergence rate and a non-normal limiting distribution. Our approach does not rely on the specification of an outcome regression model. We also investigate the doubly robust estimator for the quantile-optimal treatment regime, which can improve the estimation efficiency when a reliable outcome regression model is available (Section 1.1 of the online supplement).

Our proposed novel representation applies to a general class of policy search estimators for optimal treatment regimes defined by a general class of criteria. In particular, our approach can be applied to investigate the asymptotic distribution for the estimators of the mean-optimal treatment regimes in Zhang et al. (2012a, 2013) and fill in an important gap in the theory. The aforementioned nonstandard asymptotics will also arise when the mean-optimal criterion is used. For alternative criteria, we discuss optimal treatment regimes defined by the Gini's mean difference criterion and the weighted quantile criterion in the online supplement, where an outline of the theory and some numerical examples are provided.

It is worth noting that the nonstandard asymptotics discussed in this paper are different from the nonregular asymptotics for Q-learning estimators. The Q-learning method models the stage-specific conditional mean functions and is a popular indirect method for estimating mean-optimal treatment regimes. Consider the Q-learning method in a two-stage dynamic setting and denote the estimated parameters indexing the optimal treatment regimes for the two stages as $(\hat{\psi}_1, \hat{\psi}_2)$. The asymptotic distribution for $\hat{\psi}_2$ is standard but the asymptotic distribution for $\hat{\psi}_1$ is nonregular in the sense that it does not converge uniformly over the parameter space (Robins, 2004; Chakraborty et al., 2010; Laber et al., 2014). The asymptotic distribution of $\hat{\psi}_1$ can change abruptly from being asymptotically normal to being non-normal depending on whether a certain event occurs with probability zero or not. This happens because $\hat{\psi}_1$ is a nonsmooth function of $\hat{\psi}_2$. The results in this paper and those in the literature on Q-learning demonstrate the challenges of asymptotic theory for optimal treatment regimes estimation. In general, classical asymptotic theory is no longer applicable.

An interesting future research direction is to investigate estimating quantile-optimal treatment regimes for survival data, where the response variable is randomly censored. Censored data arise in diverse fields such as economics, medicine and sociology. For example, in a clinical trial censoring occurs when a study ends before all patients experience the event of interest. Several authors (Goldberg and Kosorok (2012); Zhao et al. (2015c); Geng et al. (2015); Jiang et al. (2016)) recently studied estimating optimal treatment regimes with survival outcomes but have not considered the quantile criterion. When censoring is heavy, it can be difficult to estimate the mean survival time accurately but it is often possible to reliably estimate the median and the lower quantiles.

References

Behncke, S., Froelich, M., and Lechner, M. (2009). Targeting labour market programmes: Results from a randomized experiment. *Swiss Journal of Economics and Statistics*, 145(3):221–268.

Bhattacharya, D. (2009). Inferring optimal peer assignment from experimental data. *Journal of the American Statistical Association*, 104(486):486–500.

Bhattacharya, D. and Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196.

Burger, D. M., Meenhorst, P. L., ten Napel, C. H., Mulder, J. W., Neef, C., Koks, C. H., Bult, A., and Beijnen, J. H. (1994). Pharmacokinetic variability of zidovudine in hiv-infected individuals: subgroup analysis and drug interactions. *AIDS*, 8(12):1683–1690.

Cai, T., Tian, L., Wong, P. H., and Wei, L. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics*, 12(2):270–282.

Chakraborty, B. and Moodie, E. E. (2013). *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer Science & Business Media.

Chakraborty, B., Murphy, S., and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, 19(3):317–343.

Chakraborty, B. and Murphy, S. A. (2014). Dynamic treatment regimes. *Annual Review of Statistics and its Application*, 1:447.

Chernozhukov, V. and Hansen, C. (2005). An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics*, 125(1):141–173.

Frölich, M. (2008). Statistical treatment choice: an application to active labor market programs. *Journal of the American Statistical Association*, 103:547–558.

Geng, Y., Zhang, H. H., and Lu, W. (2015). On optimal treatment regimes selection for mean survival time. *Statistics in medicine*, 34(7):1169–1184.

Gerber, A. S. and Green, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review*, 94:653–663.

Goldberg, Y. and Kosorok, M. R. (2012). Q-learning with censored data. *Annals of Statistics*, 40(1):529.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090.

Henderson, R., Ansell, P., and Alshibani, D. (2010). Regret-regression for optimal dynamic treatment regimes. *Biometrics*, 66(4):1192–1201.

Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701.

Hogan, J. W. and Lee, J. Y. (2004). Marginal structural quantile models for longitudinal observational studies with time-varying treatment. *Statistica Sinica*, pages 927–944.

Huang, X., Choi, S., Wang, L., and Thall, P. F. (2015). Optimization of multi-stage dynamic treatment regimes utilizing accumulated data. *Statistics in medicine*, 34(26):3424–3443.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7:443–470.

Jiang, R., Lu, W., Song, R., and Davidian, M. (2016). On estimation of optimal treatment regimes for maximizing t-year survival probability. *Journal of the Royal Statistical Society: Series B*. In Press.

Kim, J. K. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, 1:191–219.

Kosorok, M. R. and Moodie, E. E. (2016). *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine.* ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics*, 8(1):1225.

Lavori, P. W. and Dawson, R. (2000). A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society: Series A*, 163:29–38.

Linn, K. A., Laber, E. B., and Stefanski, L. A. (2015). Interactive q-learning for probabilities and quantiles. *arXiv:1407.3414*.

Loomis, L. H. and Sternberg, S. (1968). *Advanced Calculus*. Addison-Wesley, Reading, Mass.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246.

Matsouaka, R. A., Li, J., and Cai, T. (2014). Evaluating marker-guided treatment selection strategies. *Biometrics*, 70(3):489–499.

Moodie, E., Dean, N., and Sun, Y. (2014). Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 6:223–243.

Moodie, E. E., Platt, R. W., and Kramer, M. S. (2009). Estimating response-maximized decision rules with applications to breastfeeding. *Journal of the American Statistical Association*, 104:155–165.

Moodie, E. E. and Richardson, T. S. (2010). Estimating optimal dynamic regimes: Correcting bias under the null. *Scandinavian Journal of Statistics*, 37(1):126–146.

Moodie, E. E., Richardson, T. S., and Stephens, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics*, 63(2):447–455.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B*, 65(2):331–366.

Murphy, S. A. (2005a). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24(10):1455–1481.

Murphy, S. A. (2005b). A generalization error for q-learning. *Journal of Machine Learning Research*, 6:1073–1097.

Murphy, S. A. (2008). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24:1455–1481.

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472.

Orellana, L., R. A. and Robins, J. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: Main content. *The International Journal of Biostatistics*, 6.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.*, 39(2):1180–1210.

Qian, M., Nahum-Shani, I., and Murphy, S. A. (2012). Dynamic treatment regimes. In *Modern Clinical Trial Analysis*, pages 127–148. Springer.

Robins, J., Hernan, M., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality (Los Angeles, CA, 1994)*, volume 120 of *Lecture Notes in Statist.*, pages 69–117. Springer, New York.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, pages 189–326. Springer.

Robins, J.M., O. L. and Rotnitzky, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27:4678–4721.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of educational Psychology*, 66:688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics*, 6:34–58.

Rubin, D. B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association*, 81:961–962.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*, 29(4):640.

Song, R., Wang, W., Zeng, D., and Kosorok, M. (2015). Penalized q-learning for dynamic treatment regimens. *Statistica Sinica*, 25:901–920.

Staghøj, J., Svarer, M., and Rosholm, M. (2010). Choosing the best training programme: Is there a case for statistical treatment rules? *Oxford Bulletin of Economics and Statistics*, 72:172–201.

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81.

Tao, Y. and Wang, L. (2017). Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics*, 73(1):145–155.

Tetenov, A. (2012). Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, 166(1):157–165.

Thall, P. F., Sung, H.-G., and Estey, E. H. (2011). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *Journal of the American Statistical Association*, 97:29–39.

van der Laan, M. J., Petersen, M. L., and Joffe, M. M. (2005). History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *Journal of Biostatistics*, 1.

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York.

Wallace, M. P. and Moodie, E. E. (2014). Personalizing medicine: a review of adaptive treatment strategies. *Pharmacoepidemiology and Drug Safety*, 23(6):580–585.

Watkins, C. and Dayan, P. (1992). Q-learning. *Maching Learning*, 8:279–292.

Wunsch, C. (2013). Optimal use of labor market policies: the role of job search assistance. *Review of Economics and Statistics*, 95(3):1030–1045.

Zhang, B., Tsiatis, A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100:681–694.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012a). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018.

Zhang, Z., Chen, Z., Troendle, J. F., and Zhang, J. (2012b). Causal inference on quantiles with an obstetric application. *Biometrics*, 68:697–706.

Zhao, Y., Zeng, D., Laber, E., Song, R., Yuan, M., and Kosorok, M. (2015a). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102:151–168.

Zhao, Y. Q., Zeng, D., Laber, E. B., and Kosorok, M. R. (2015b). New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110:583–598.

Zhao, Y.-Q., Zeng, D., Laber, E. B., Song, R., Yuan, M., and Kosorok, M. R. (2015c). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168.

Zhao, Y. Q., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.

Zhou, Y., Wang, L., Sherwood, B., and Song, R. (2017). quantoptr: Algorithms for quantile- and mean-optimal treatment regimes. `https://CRAN.R-project.org/package=quantoptr`.

Appendix: Technical Proofs

We provide below the proofs of Lemma 2 and Theorem 1. The proofs of Lemma 1, Theorem 2, and derivation of the theory for Section 6.2 are given in the online supplement.

Proof of Lemma 2. (1) Note that $g(\cdot,\beta,m) = \big[AI(X^T\beta>0)+(1-A)I(X^T\beta\le 0)\big]I(Y-m>0)$. The classes $\{I(X^T\beta>0):\beta\in\mathbb{B}\}$ and $\{I(Y-m>0):m\in\mathbb{R}\}$ are both VC subgraph classes and hence bounded Donsker classes. Therefore, the class $\{g(\cdot,\beta,m):\beta\in\mathbb{B},m\in\mathbb{R}\}$ is Donsker (van der Vaart and Wellner (1996)). We thus have

$$\sup_{\beta\in\mathbb{B},m\in\mathbb{R}}\big|P_ng(\cdot,\beta,m)-Pg(\cdot,\beta,m)\big| = O_p(n^{-1/2}). \tag{8}$$

We denote the supremum at the left side of the above expression as Δ_n. For any given β, $Pg(\cdot, \beta, m)$ is a decreasing function of m. Hence the assumption about the density ensures that there exists a constant $\kappa_1 > 0$ such that $\sup_{\beta \in \mathbb{B}} Pg(\cdot, \beta, m_0 + \epsilon) < \frac{1-\tau}{2} - \kappa_1 \epsilon$, for each small enough $\epsilon > 0$. Taking $\epsilon = \Delta_n/\kappa_1$, for all n sufficiently large, it follows from (8) that $\sup_{\beta \in \mathbb{B}} P_n g(\cdot, \beta, m_0 + \Delta_n/\kappa_1) < \Delta_n + \frac{1-\tau}{2} - \kappa_1 \frac{\Delta_n}{\kappa_1} = \frac{1-\tau}{2}$. This implies $\widehat{m}_n < m_0 + \Delta_n/\kappa_1$ for all n sufficiently large. Similarly, there exists a constant $\kappa_2 > 0$ such that $\sup_{\beta \in \mathbb{B}} Pg(\cdot, \beta, m_0 - \epsilon) \geq \frac{1-\tau}{2} + \kappa_2 \epsilon$, for all small enough $\epsilon > 0$. If follows that $\sup_{\beta \in \mathbb{B}} P_n g(\cdot, \beta, m_0 - \Delta_n/\kappa_2) \geq -\Delta_n + \frac{1-\tau}{2} + \kappa_2 \frac{\Delta_n}{\kappa_2} = (1-\tau)/2$ for all n sufficiently large. This implies $\widehat{m}_n \geq m_0 - \Delta_n/\kappa_2$ for all n sufficiently large. Since $\Delta_n = O_p(n^{-1/2})$, we have $\widehat{m}_n = m_0 + O_p(n^{-1/2})$.

(2) Observing (i) $\widehat{\beta}_n = \underset{\beta \in \mathbb{B}}{\operatorname{argmax}} P_n g(\cdot, \beta, \widehat{m}_n)$, (ii) $\beta = \beta_0$ uniquely maximizes $Pg(\cdot, \beta, m_0)$ and (iii) $\sup_{\beta \in \mathbb{B}} \left| P_n g(\cdot, \beta, \widehat{m}_n) - Pg(\cdot, \beta, m_0) \right| = o_p(1)$, we conclude that $\widehat{\beta}$ is consistent for β_0 by applying standard arguments of the M estimation theory (simple modification of Theorem 5.7 in van der Vaart (1998)). Next, we will show $\widehat{\beta}_n - \beta_0 = O_p(n^{-1/3})$.

Let $\theta = (\beta^T, \delta)^T$, where $\delta = m - m_0$, and $h(\cdot, \beta, \delta) = C(\beta) I\{Y - m_0 - \delta > 0\} - C(\beta_0) I\{Y - m_0 - \delta > 0\}$. By definition, $\widehat{\beta}_n = \underset{\beta \in \mathbb{B}}{\operatorname{argmax}} P_n h(\cdot, \beta, \widehat{m}_n - m_0)$. We will consider a Taylor expansion of $Ph(\cdot, \beta, \delta)$ around $\theta_0 = (\beta_0^T, 0)^T$. Note that $h(\cdot, \beta_0, 0) = 0$ and that

$$
\begin{aligned}
& E\big[C(\beta) I\{Y - m_0 - \delta > 0\}\big] \\
= \; & \frac{1}{2} E\big\{ I(X^T\beta > 0) I(Y - m_0 - \delta > 0) \big| A = 1 \big\} + \frac{1}{2} E\big\{ I(X^T\beta \leq 0) I(Y - m_0 - \delta > 0) \big| A = 0 \big\} \\
= \; & \frac{1}{2} E\big\{ I(X^T\beta > 0) S_{1,X}(m_0 + \delta) \big\} + \frac{1}{2} E\big\{ I(X^T\beta \leq 0) S_{0,X}(m_0 + \delta) \big\} \\
= \; & \frac{1}{2} E\big\{ I(X^T\beta > 0)(S_{1,X}(m_0 + \delta) - S_{0,X}(m_0 + \delta)) \big\} + \frac{1}{2} E\big\{ S_{0,X}(m_0 + \delta) \big\},
\end{aligned}
$$

where $S_{1,X}(\cdot)$ and $S_{0,X}(\cdot)$ are the conditional survival functions of $Y^*(1)$ and $Y^*(0)$ given X, respectively. Let $q(X, \delta) = S_{1,X}(m_0 + \delta) - S_{0,X}(m_0 + \delta)$, then

$$
E(h(\cdot, \beta, \delta)) = \frac{1}{2} E\big\{ \big(I(X^T\beta > 0) - I(X^T\beta_0 > 0) \big) q(X, \delta) \big\}.
$$

It is easy to see $\frac{\partial}{\partial\delta}E(h(\cdot,\beta,\delta))\big|_{\beta=\beta_0,\delta=0} = 0$ and $\frac{\partial^2}{\partial\delta^2}E(h(\cdot,\beta,\delta))\big|_{\beta=\beta_0,\delta=0} = 0$. Note that the transformation $T_\beta = (I - ||\beta||^{-2}\beta\beta^T)(I - \beta_0\beta_0^T) + ||\beta||^{-1}\beta\beta_0^T$, where I denotes the identity matrix, maps the region $A = \{X^T\beta_0 > 0\}$ onto $A(\beta) = \{X^T\beta > 0\}$, taking ∂A to $\partial A(\beta)$. The surface measure σ_β on $\partial A(\beta)$ has the constant density $\rho_\beta(X) = \beta^T\beta_0/||\beta||$ with respect to the image of the surface measure $\sigma = \sigma_{\beta_0}$ under T_β. The outward pointing unit vector normal to $A(\beta)$ is the standardized vector $-\beta/||\beta||$ and along ∂A the derivative $(\partial/\partial\beta)T_\beta(X)$ reduces to $-||\beta||^{-2}[\beta X^T + (\beta^T X)I]$. Using the result from Section 10.7 of Loomis and Sternberg (1968) on derivatives as surface integrals, we have

$$\frac{\partial}{\partial\beta^T}E(h(\cdot,\beta,\delta)) = \frac{1}{2}||\beta||^{-2}\beta^T\beta_0(I + ||\beta||^{-2}\beta\beta^T)\int I\{X^T\beta_0 = 0\}q(T_\beta(X),\delta)f(T_\beta(X))X d\sigma.$$

Note that we have $\frac{\partial}{\partial\beta}E(h(\cdot,\beta,\delta))\big|_{\beta=\beta_0,\delta=0} = 0$ because $E(h(\cdot,\beta,0))$ is maximized at $\beta = \beta_0$. Combining with the observation that $T_{\beta_0}(X) = X$ along $\{X^T\beta_0 = 0\}$, we have $\int I\{X^T\beta_0 = 0\}l(X,0)f(X)X d\sigma = 0$. Using this and the fact $||\beta_0|| = 1$, we have

$$\frac{\partial^2}{\partial\beta\partial\beta^T}E(h(\cdot,\beta,\delta))\big|_{\beta=\beta_0,\delta=0} = -\frac{1}{2}\int I\{X^T\beta_0 = 0\}(f(X)\dot{q}(X,0) + q(X,0)\dot{f}(X))^T\beta_0 XX^T d\sigma,$$

where $\dot{q}(X,0)$ and $\dot{f}(X)$ denote the gradients with respect to X. Also,

$$\frac{\partial^2}{\partial\beta^T\partial\delta}E(h(\cdot,\beta,\delta))\big|_{\beta=\beta_0,\delta=0} = \frac{1}{2}\int I\{X^T\beta_0 = 0\}(s_{1,X}(m_0) - s_{0,X}(m_0))f(X)X d\sigma,$$

where $s_{1,X}$ and $s_{0,X}$ are the derivatives of $S_{1,X}$ and $S_{0,X}$ with respect to δ, respectively. We write

$$V = -\frac{\partial^2}{\partial\beta\partial\beta^T}E(h(\cdot,\beta,\delta))\big|_{\beta=\beta_0,\delta=0} \tag{9}$$

and $a_1 = \frac{\partial^2}{\partial\beta^T\partial\delta}E(h(\cdot,\beta,\delta))\big|_{\beta=\beta_0,\delta=0}$, then the Taylor expansion of $Ph(\cdot,\beta,\delta)$ around $(\beta_0,0)$

has the form

$$Ph(\cdot,\beta,\delta) \;=\; -\frac{1}{2}(\beta-\beta_0)^T V(\beta-\beta_0) + a_1^T(\beta-\beta_0)\delta + o(||\beta-\beta_0||^2) + o(\delta^2). \quad (10)$$

For a given positive constant R, let $H_R = \sup_{||\theta-\theta_0||\le R} |h(\cdot,\beta,\delta)|$. We observe that $h(\cdot,\beta,\delta)$ is nonzero if and only if $C(\beta)$ and $C(\beta_0)$ take different values. Hence, $H_R \le \sup_{||\theta-\theta_0||\le R} \{I(X^T\beta > 0 \ge X^T\beta_0) + I(X^T\beta_0 > 0 \ge X^T\beta)\}$. The envelope function H_R is bounded by an indicator function of a pair of multidimensional wedge shaped regions, each subtending an angle of order $O(R)$, from which we deduce that $E(H_R^2) = O(R)$. The conditions of Lemma 4.1 of Kim and Pollard (1990) are satisfied. Hence, for each fixed $\epsilon > 0$, uniformly for $||\theta-\theta_0|| \le R$, $P_n h(\cdot,\beta,\delta) \le Ph(\cdot,\beta,\delta) + \epsilon(||\beta-\beta_0||^2 + \delta^2) + O_p(n^{-2/3})$. Combining with the upper bound in (10), we have $P_n h(\cdot,\beta,\delta) \le -\big(\frac{1}{2}\lambda_{min}(V) - \epsilon\big)||\beta-\beta_0||^2 + ||a_1||||\beta-\beta_0|||\delta| + (\epsilon + o(1))\delta^2 + O_p(n^{-2/3})$, where $\lambda_{min}(V)$ denotes the smallest eigenvalue of V. Choosing $\epsilon = \lambda_{min}(V)/4$, we derive that

$$\begin{aligned} 0 \;&=\; P_n h(\cdot,\beta_0,\widehat{m}_n - m_0) \le P_n h(\cdot,\widehat{\beta}_n,\widehat{m}_n - m_0) \\ &\le\; -\frac{1}{4}\lambda_{min}(V)||\widehat{\beta}_n - \beta_0||^2 + O_p(n^{-1/2})||\widehat{\beta}_n - \beta_0|| + O_p(n^{-2/3}). \end{aligned}$$

Completing the square in $||\widehat{\beta}_n - \beta_0||$, we derive that $||\widehat{\beta}_n - \beta_0|| = O_p(n^{-1/3})$.

Next, we show that $\widehat{\beta}_n$ nearly maximizes $P_n h(\cdot,\beta,0)$. A similar argument as above shows that $P|h(\cdot,\theta_1) - h(\cdot,\theta_2)| = O(||\theta_1 - \theta_2||)$ for θ_1, θ_2 near θ_0. It follows from Lemma 4.6 of Kim and Pollard (1990) that the process $J_n(\cdot,\alpha,\gamma) = n^{2/3}(P_n - P)h(\cdot,\beta_0 + \alpha n^{-1/3}, \gamma n^{-1/3})$ satisfies the stochastic equicontinuity condition of Theorem 2.3 of Kim and Pollard (1990). Since $n^{1/3}(\widehat{m}_n - m_0) = o_p(1)$, this implies that for β uniformly in a $O(n^{-1/3})$ neighborhood of β_0, $J_n(\cdot, n^{1/3}(\beta-\beta_0), n^{1/3}(\widehat{m}_n - m_0)) - J_n(\cdot, n^{1/3}(\beta-\beta_0), 0) = o_p(1)$. That is, $P_n h(\cdot,\beta,\widehat{m}_n - m_0) = P_n h(\cdot,\beta,0) + Ph(\cdot,\beta,\widehat{m}_n - m_0) - Ph(\cdot,\beta,0) + o_p(n^{-2/3})$, uniformly over an $O_p(n^{-1/3})$ neighborhood of β_0. Within such a neighborhood, Taylor expansion similarly as before

shows that $Ph(\cdot,\beta,\widehat{m}_n - m_0) - Ph(\cdot,\beta,0) = o_p(n^{-2/3})$. Suppose $\widetilde{\beta}_n = \underset{\beta\in\mathbb{B}}{\text{argmax}}\ P_n h(\cdot,\beta,0)$. An analysis similar to that for $\widehat{\beta}_n$ shows that $\widetilde{\beta}_n = O_p(n^{-1/3})$. Hence,

$$\begin{aligned} P_n h(\cdot,\widehat{\beta}_n,0) &= P_n h(\cdot,\widehat{\beta}_n,\widehat{m}_n - m_0) - o_p(n^{-2/3}) \geq P_n h(\cdot,\widetilde{\beta}_n,\widehat{m}_n - m_0) - o_p(n^{-2/3}) \\ &= P_n h(\cdot,\widetilde{\beta}_n,0) - o_p(n^{-2/3}), \end{aligned}$$

where the inequality follows because $\widehat{\beta}_n = \underset{\beta\in\mathbb{B}}{\text{argmax}}\ P_n h(\cdot,\beta,\widehat{m}_n - m_0)$. Therefore, $P_n h(\cdot,\widehat{\beta}_n,0) \geq \sup_{\beta\in\mathbb{B}} P_n h(\cdot,\beta,0) - o_p(n^{-2/3})$. □

Proof of Theorem 1. Following Lemma 2(2), to find the asymptotic distribution of $n^{1/3}(\widehat{\beta}_n - \beta_0)$, it suffices to apply the main theorem of Kim and Pollard (1990) to the one parameter process $\{h(\cdot,\beta,0) : \beta \in \mathbb{B}\}$. Recall that $h(\cdot,\beta,0) = C(\beta)I\{Y > m_0\} - C(\beta_0)I\{Y > m_0\}$. In the following, we will verify conditions (iv) and (v) of the main theorem of Kim and Pollard (1990). Other conditions of the theorem are relatively easier and can be checked using similar techniques as those in the proof of Lemma 2.

For condition (iv), it can be shown that $\frac{\partial^2}{\partial\beta\partial\beta^T} E(h(\cdot,\beta,0))\big|_{\beta=\beta_0} = -V$, where V is defined in (9) in the proof of Lemma 2. Next, we calculate the kernel function in condition (v). Similarly as in the calculation in the proof of Lemma 2, for each C_1, C_2 in R^l, and $t > 0$,

$$\begin{aligned} & tP\Big|h\Big(\cdot,\beta_0 + \frac{C_1}{t},0\Big) - h\Big(\cdot,\beta_0 + \frac{C_2}{t},0\Big)\Big|^2 \\ = \ & tP\Big\{\big|C(\beta_0 + C_1/t) - C(\beta_0 + C_2/t)\big| I(Y > m_0)\Big\} \\ = \ & \frac{1}{2}tP\Big\{\big|I(X^T(\beta_0 + C_1/t) > 0) - I(X^T(\beta_0 + C_2/t) > 0)\big| I(Y^*(1) > m_0)\Big\} \\ & + \frac{1}{2}tP\Big\{\big|I(X^T(\beta_0 + C_1/t) \leq 0) - I(X^T(\beta_0 + C_2/t) \leq 0)\big| I(Y^*(0) > m_0)\Big\} \\ = \ & tP\big\{(S_{1,X}(m_0) + S_{0,X}(m_0))\big|I(X^T(\beta_0 + C_1/t) > 0) - I(X^T(\beta_0 + C_2/t) > 0)\big|\big\}. \end{aligned}$$

To evaluate the above expression, we make use of the local coordinates (Example 6.4 of Kim and Pollard (1990)), for which we define $\beta(\tau) = \sqrt{1 - ||\tau||^2}\beta_0 + \tau$, where τ is orthogonal to

β_0 and ranges over a neighborhood of the origin. It is noted that as the parameter space is on the sphere ($||\beta_0|| = 1$, $||\beta|| = 1$), such a decomposition can be obtained by taking $\tau = \tau(\beta) = T_0\beta$, where $T_0 = I - \beta_0\beta_0^T$. Then we can write $\beta = (\beta_0^T\beta)\beta_0 + T_0\beta$ such that $\beta_0^T\beta = \sqrt{1 - ||\tau||^2}$ and $\beta_0^T T_0\beta = 0$. Also, $\tau(\beta_0 + C_1/t) = T_0C_1/t$, $\tau(\beta_0 + C_2/t) = T_0C_2/t$. Similarly, we can decompose X as $X = r\beta_0 + Z$ for some random variable r and random vector Z, with Z being orthogonal to β_0. Let $C_k^* = T_0C_k \in T_0$, $k = 1, 2$, then $X^T(\beta_0+C_1/t) = (r\beta_0 + Z)^T\big(\sqrt{1 - ||C_1^*||^2/t^2}\beta_0 + C_1^*/t\big) = r\sqrt{1 - ||C_1^*||^2/t^2} + Z^TC_1^*/t$. Let $p(\cdot, \cdot)$ be the joint density function of (r, Z), which can be deduced from the density of X, With a change of variable $w = tr$, $tP\big\{(S_{1,X}(m_0) + S_{0,X}(m_0))\big|I(X^T(\beta_0 + C_1/t) > 0) - I(X^T(\beta_0 + C_2/t) > 0)\big|\big\}$ is equal to

$$\begin{aligned}
&\iint I\big\{ - Z^TC_2^*(1 - ||C_2^*||^2/t^2)^{-1/2} > w \geq -Z^TC_1^*(1 - ||C_1^*||^2/t^2)^{-1/2}\big\} \\
&\qquad (S_{1,\frac{w}{t}\beta_0+Z}(m_0) + S_{0,\frac{w}{t}\beta_0+Z}(m_0))p(w/t, Z)dwdZ \\
&+ \iint I\big\{ - Z^TC_1^*(1 - ||C_1^*||^2/t^2)^{-1/2} > w \geq -Z^TC_2^*(1 - ||C_2^*||^2/t^2)^{-1/2}\big\} \\
&\qquad \big(S_{1,\frac{w}{t}\beta_0+Z}(m_0) + S_{0,\frac{w}{t}\beta_0+Z}(m_0)\big)p(w/t, Z)dwdZ.
\end{aligned}$$

Integrating over w and letting $t \to \infty$ to get $\lim_{t\to\infty} tP\big|h(\cdot, \beta_0+C_1/t, 0)-h(\cdot, \beta_0+C_2/t, 0)\big|^2 = \int |Z^T(C_1^*-C_2^*)|\big(S_{1,Z}(m_0)+S_{0,Z}(m_0)\big)p(0, Z)dZ = \int |Z^T(C_1-C_2)|\big(S_{1,Z}(m_0)+S_{0,Z}(m_0)\big)p(0, Z)dZ$. We denote this limit as $L(C_1-C_2)$. Using the identity $2xy = x^2+y^2-(x-y)^2$, we deduce that the limiting covariance kernel function can be written as $K(C_1, C_2) = \lim_{t\to\infty} tP\big\{h(\cdot, \beta_0 + C_1/t, 0)h(\cdot, \beta_0+C_2/t/t, 0)\big\} = \lim_{t\to\infty} \frac{1}{2}tP\big|h(\cdot, \beta_0+C_1/t, 0)-h(\cdot, \beta_0, 0)\big|^2+\lim_{t\to\infty} \frac{1}{2}tP\big|h(\cdot, \beta_0+C_2/t, 0)-h(\cdot, \beta_0, 0)\big|^2-\lim_{t\to\infty} \frac{1}{2} \lim_{t\to\infty} tP\big|h(\cdot, \beta_0+C_1/t, 0)-h(\cdot, \beta_0+C_2/t, 0)\big|^2 = \frac{1}{2}\big(L(C_1)+L(C_2) - L(C_1 - C_2)\big)$. The asymptotic distribution of $n^{1/3}(\widehat{\beta}_n - \beta_0)$ then follows by applying the main theorem of Kim and Pollard (1990) $\square$

Online Supplement to JASA-T&M-2015-0234R2

"Quantile-Optimal Treatment Regimes"

Lan Wang, Yu Zhou, Rui Song and Ben Sherwood

1 Methodology extensions

1.1 Doubly robust estimation for observational study

For the mean criterion, Zhang et al. (2012) investigated an alternative approach based on an augmented estimator of the mean of the potential outcome. Their estimator relies on a model for the conditional mean $E(Y|A,X)$. The iterative expectation formula naturally connects the conditional mean to the marginal mean. The alternative estimator enjoys the double robustness property, that is, it is consistent if either the propensity score model or the conditional mean regression model is correctly specified.

Doubly robust estimation for the quantile criterion is challenging as we do not have such a natural link between a conditional quantile and a marginal quantile. We proceed as follows. Assume that $Q_\tau(Y^*(1)|X) = g(X,\beta_1(\tau))$ and $Q_\tau(Y^*(0)|X) =$

[1]Lan Wang is Professor and Yu Zhou is graduate student, School of Statistics, University of Minnesota, Minneapolis, MN 55455. Emails: wangx346@umn.edu and zhou0269@umn.edu. Rui Song is Associate Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. Email: rsong@ncsu.edu. Ben Sherwood is Assistant Professor, School of Business, University of Kansas. Email: ben.sherwood@ku.edu. Dr. Sherwood's work was done when he was a graduate student at University of Minnesota. Wang's research is partly supported by NSF DMS-1512267 and DMS-1712706. Song's research is partly supported by NSF DMS-1555244 and NCI P01 CA142538. We thank the Co-Editor Nicholas Jewell, the AE and two anonymous referees for their constructive comments which help us significantly improve the paper. We also thank Dr. Shannon Holloway at North Carolina State University for proofreading the paper and providing many helpful comments.

$g(X, \beta_0(\tau))$, for all $\tau \in (0,1)$, where $g(\cdot, \beta_1(\tau))$ and $g(\cdot, \beta_0(\tau))$ are parametrically specified model for the τth conditional quantile for the potential outcomes $Y^*(1)$ and $Y^*(0)$, respectively. Let $\{u_{11}, \ldots, u_{1n}\}$ and $\{u_{01}, \ldots, u_{0n}\}$ be independent random samples from Uniform$(0,1)$ distribution. If we generate random responses by $Y_i^{**} = g(X_i, \beta_1(u_{1i}))d(X_i, \beta) + g(X_i, \beta_0((u_{0i}))(1 - d(X_i, \beta))$, $i = 1, \ldots, n$, then we obtain a random sample from the distribution of $Y^*(d_\beta)$, as motivated by the random coefficient interpretation of quantile regression (Section 2.6, Koenker, 2005).

In practice, we replace $\beta_1(\tau)$ and $\beta_0(\tau)$ by their estimators, which are obtained by separate conditional quantile regression on the observations from the group $A = 1$ and the group $A = 0$, respectively. Let the correspondingly generated outcomes be denoted by $\widehat{Y}_i^{**}$, $i = 1, \ldots, n$. We estimate the τth quantile of $Y^*(d_\beta)$ by

$$\tilde{Q}_\tau(\beta) = \underset{a}{\text{argmin}} n^{-1} \sum_{i=1}^{n} \Big[\frac{C_i(\beta)}{\hat{\pi}_c(X_i, \beta)} \rho_\tau(Y_i - a) + \big(1 - \frac{C_i(\beta)}{\hat{\pi}_c(X_i, \beta)}\big) \rho_\tau(\widehat{Y}_i^{**} - a) \Big],$$

where $\hat{\pi}_c(X_i, \beta)$ is the estimated propensity score as described in Remark 2 of Section 3.3 of the main paper. This estimator can be shown to enjoy the double robustness property. The estimator of the parameter indexing the optimal treatment regime is defined the same as before. A numerical example is provided in Section 3.3 of this online supplement.

To obtain the doubly robust estimator, we can implement an algorithm similar as that described in Section 3.4 of the main paper. The main change is to apply the following steps to modify the function quant_est to include the model-based simulated outcome $\widehat{Y}^{**}$.

1. We generate two independent random samples from the uniform distribution on the interval (0,1): $\{u_{11}, \ldots, u_{1n}\}$ and $\{u_{01}, \ldots, u_{0n}\}$.

2. Using observations with $A_i = 1$, we estimate the conditional quantile $Q_\tau(Y^*(1)|X) =$

$g(X, \beta_1(\tau))$, for τ in a fine grid of (0,1). This yields $\widehat{\beta}_1(\tau)$, $\tau \in (0,1)$.

3. Using observations with $A_i = 0$, we estimate the conditional quantile $Q_\tau(Y^*(0)|X) = g(X, \beta_0(\tau))$, for τ in a fine grid of (0,1). This yields $\widehat{\beta}_0(\tau)$, $\tau \in (0,1)$.

4. Calculate $\widehat{Y}_i^{**} = g(X_i, \widehat{\beta}_1(u_{1i}))d(X_i, \beta) + g(X_i, \widehat{\beta}_0((u_{0i}))(1-d(X_i, \beta))$, $i = 1, \ldots, n$.

5. The objective function in the function quant_est (Listing 1 of the main paper.) is now modified to $n^{-1}\sum_{i=1}^{n}\left[\frac{C_i(\beta)}{\hat{\pi}_c(X_i,\beta)}\rho_\tau(Y_i - a) + \big(1 - \frac{C_i(\beta)}{\hat{\pi}_c(X_i,\beta)}\big)\rho_\tau(\widehat{Y}_i^{**} - a)\right]$.

1.2 Optimal treatment regimes with respect to alternative criteria

The approach introduced in this paper can be adapted to estimating optimal treatment regimes defined by other criteria. For example, economists are sometimes interested in finding the policy that leads to the smallest dispersion in the income distribution. Motivated by the helpful suggestions from an anonymous referee, we first consider estimating the optimal treatment regime with the goal to minimize Gini's mean difference (Gini, 1912), a robust measure of dispersion.

Given a treatment regime d_β, let $G(Y^*(d_\beta)) = -E\big|Y_1^*(d_\beta) - Y_2^*(d_\beta)\big|$, where $Y_1^*(d_\beta)$ and $Y_2^*(d_\beta)$ are independent copies of the potential outcome $Y^*(d_\beta)$. The optimal treatment regime minimizing Gini's mean difference is defined as $\arg max_{d\in\mathbb{D}} G(Y^*(d_\beta))$. For data from a randomized trial, we can consistently estimate $G(Y^*(d_\beta))$ using the following second-order U-statistic

$$\widehat{G}(\beta) = -\frac{2}{n(n-1)}\sum_{1\le i<j\le n} 4C_i(\beta)C_j(\beta)|Y_i - Y_j|.$$

The estimated parameter indexing the optimal treatment regime is $\widehat{\beta}_n = \underset{\beta\in\mathbb{B}}{\text{argmax}}\, \widehat{G}(\beta)$. Let $g(Z_i, Z_j, \xi, m) = 4C_i(\beta)C_j(\beta)(-|Y_i - Y_j| - m)$, $\widehat{m}_n = \sup\{m : \sup_{\beta\in\mathbb{B}} U_n g(\cdot, \cdot, \xi, m) \ge$

$0\}$, $m_0 = \sup\{m : \sup_{\beta\in\mathbb{B}} Pg(\cdot,\cdot,\xi,m) \geq 0\}$, $\beta_0 = \underset{\beta\in\mathbb{B}}{\text{argmax}}\ Pg(\cdot,\cdot,\xi,m_0)$, where $Z_i = \{X_i, Y_i\}$, $U_n g(\cdot,\cdot,\beta,m) = \frac{2}{n(n-1)}\sum_{1\leq i<j\leq n} g(Z_i, Z_j, \xi, m)$. Then we have $\widehat{\beta}_n = \underset{\beta\in\mathbb{B}}{\text{argmax}}\ U_n g(\cdot,\cdot,\beta,\widehat{m}_n)$. The derivation of the asymptotic distribution of $\widehat{\beta}_n$ follows similar ideas as that for quantile-optimal treatment regime but involves two new technical challenges. First, unlike for the mean- or quantile criterion, the objective function of Gini's mean difference is a second-order U-statistics. Hence, the derivation involves more delicate application of U-process uniform convergence theory. Second, the calculation of the V matrix in the quadratic drift of the Gaussian process involves more complex derivatives of surface integral. In Section 2.3 of this online supplement, we showed that the limiting distribution of $n^{1/3}(\widehat{\xi}_n - \xi_0)$ is that of the maximizer of a centered Gaussian process with a quadratic drift. A simulation example is provided in Section 3.4 of this online supplement.

The referee also suggested that in practice we are sometimes interested in finding the optimal treatment regime that improves the general performance with respect to the lower tail of the potential outcome distribution, for which a weighted quantile criterion could be useful. Let $\boldsymbol{\tau} = (\tau_1, ..., \tau_K)^T$, where $0 < \tau_1 < \tau_2 < ... < \tau_K < 1$, denote a sequence of quantile levels. Let $\mathbf{w} = (w_1, ..., w_K)^T$ be a vector of positive weights such that $\sum_{k=1}^K w_k = 1$. For a given treatment regime $d(X)$, the weighted quantile of the distribution of the potential outcome $Y^*(d)$ is defined as $Q^{\mathbf{w}}(Y^*(d)) = \arg\ \underset{a}{\min}\ \sum_{k=1}^K w_k E(\rho_{\tau_k}(Y_i - a))$, where $\rho_{\tau_k}(\cdot)$ is the τ_kth quantile loss function (Section 5.5, Koenker, 2005). The weighted-quantile optimal treatment regime is $\arg max_{d\in\mathbb{D}} Q^{\mathbf{w}}(Y^*(d))$.

Consider a randomized trial and the induced missing data framework introduced in the main paper. We estimate $Q^{\mathbf{w}}(Y^*(d))$ by

$$\widehat{Q}^{\mathbf{w}}(\beta) = \arg\ \underset{a}{\min}\ n^{-1}\sum_{k=1}^{K} w_k \sum_{i=1}^{n} C_i(\beta)\,\rho_{\tau_k}(Y_i - a),$$

where $C_i(\beta)$ is defined in Section 3.1. We note that $\widehat{Q}^{\mathbf{w}}(\beta)$ can be obtained by applying the `rq.fit.hogg` function from the `quantreg` package in R. The parameter indexing the weighted-quantile optimal treatment regime is estimated by $\widehat{\beta}_n^{\mathbf{w}} = \arg\max_{\beta\in\mathbb{B}} \widehat{Q}^{\mathbf{w}}(\beta)$. Similarly, for observational studies, we estimate the parameter indexing the weighted-quantile optimal treatment regime by $\arg\max_{\beta\in\mathbb{B}}\big[\arg\min_a n^{-1}\sum_{k=1}^K w_k \sum_{i=1}^n \frac{C_i(\beta)}{\hat{\pi}_c(X_i,\beta)}\rho_{\tau_k}\big(Y_i - a\big)\big]$, where $\hat{\pi}_c(X,\beta)$ is the estimated propensity score. The theory of Section 3 of the main paper carries through to this case with only some minor changes. A numerical example is given in Section 3.5 of this online supplement.

2 Additional theoretical results

2.1 Proof of Lemma 1 of the main paper

Proof. For any given $\beta \in \mathbb{B}$, $n^{-1}\sum_{i=1}^n C_i(\beta)\rho_\tau(Y_i - a)$ is a convex function of a. By the law of large numbers and the convexity, $n^{-1}\sum_{i=1}^n C_i(\beta)\rho_\tau(Y_i - a)$ converges uniformly in a to $E(C_i(\beta)\rho_\tau(Y_i - a))$ in probability. For a randomized trial,

$$\begin{aligned} & E\big[C_i(\beta)\rho_\tau(Y_i - a)\big] \\ = \;& E[C_i(\beta)\rho_\tau(Y_i^*(d_\beta) - a)] \\ = \;& E\Big\{E[C_i(\beta)|X_i]E[\rho_\tau(Y_i^*(d_\beta) - a)|X_i]\Big\} \\ = \;& \frac{1}{2}E\big\{E[\rho_\tau(Y_i^*(d_\beta) - a)|X_i]\big\} = \frac{1}{2}E\big[\rho_\tau(Y_i^*(d_\beta) - a)\big], \end{aligned}$$

where the first equality follows because $Y_i = Y_i^*(d_\beta)$ when $C_i(\beta) = 1$; the second equality applies the iterative expectation formula and the conditional independence of $Y_i^*(d_\beta)$ and $C_i(\beta)$ given X_i; the third equality uses the fact $E[C_i(\beta)|X_i] = 0.5$ for a randomized trial. Note that $E(\rho_\tau(Y^*(d_\beta) - a))$ is continuous and uniquely minimized by $a = Q_\tau(Y^*(d_\beta))$. Hence, the consistency follows from the standard M-estimation

theory. □

2.2 Derivation of the theory for quantile-optimal dynamic treatment regimes

We assume the following regularity conditions.

(C1*) The potential outcomes $Y^*(1,1)$, $Y^*(1,0)$, $Y^*(0,1)$, $Y^*(0,0)$ have continuous distributions with bounded, continuously differentiable density functions.

(C2*) The population parameter $\xi_0 = (\beta_0^T, \gamma_0^T)^T$ that indexes the optimal dynamic treatment regime satisfies $||\beta_0|| = 1$ and $||\gamma_0|| = 1$, and is unique and an interior point of the compact set $\mathbb{B} = \mathbb{B}_1 \times \mathbb{B}_2$.

(C3*) X has a continuously differentiable density function $f(\cdot)$. The angular components of X, considered as a random element of the unit sphere $\mathbb{S}$ in $\mathbb{R}^l$, has a bounded, continuous density with respect to surface measure on $\mathbb{S}$.

(C4*) The matrix $V^* = -\frac{\partial^2}{\partial\xi\partial\xi^T} E(h(\cdot,\xi,\delta))\big|_{\xi=\xi_0,\delta=0}$ is positive definite, where $h(\cdot,\xi,\delta) = \frac{I(C_\xi=\infty)}{\pi(\xi)} I(Y - m_0 - \delta > 0) - \frac{I(C_{\xi_0}=\infty)}{\pi(\xi_0)} I(Y - m_0 - \delta > 0)$.

Lemma 2.1. *Assume condition (C1*) is satisfied. Then $\widehat{Q}_\tau(\xi)$ is a consistent estimate of the τth quantile of $Y^*(d_\xi)$.*

Proof. For any given ξ, $n^{-1}\sum_{i=1}^n \frac{I(C_{\xi,i}=\infty)}{P_1(H_{1i},\beta)P_2(H_{2i},\gamma)}\rho_\tau(Y_i - a)$ is a convex function of a. By the law of large numbers and the convexity, $n^{-1}\sum_{i=1}^n \frac{I(C_{\xi,i}=\infty)}{P_1(H_{1i},\beta)P_2(H_{2i},\gamma)}\rho_\tau(Y_i - a)$ converges uniformly in a to $E\Big[\frac{I(C_\xi=\infty)}{P_1(H_1,\beta)P_2(H_2,\gamma)}\rho_\tau(Y - a)\Big]$ in probability. Recall that

$C_\xi = \infty$ if and only if $A_1 = d_\beta(X_1)$ and $A_2 = d_\gamma(H_2(d_\beta))$. We have

$$\begin{aligned}
& E\Big[\frac{I(C_\xi = \infty)}{P_1(H_1,\beta)P_2(H_2,\gamma)}\rho_\tau(Y-a)\Big] \\
= \; & E_{X_1, X_2^*(d_\beta)} E\Big[\frac{I(A_1 = d_\beta, A_2 = d_\gamma)}{P_1(H_1,\beta)P_2(H_2,\gamma)}\rho_\tau(Y^*(d_\xi)-a)\Big| X_1, X_2^*(d_\beta(X_1))\Big] \\
= \; & E_{X_1, X_2^*(d_\beta)} E\Big[\rho_\tau(Y^*(d_\xi)-a)\Big| X_1, X_2^*(d_\beta(X_1))\Big] \\
= \; & E\big[\rho_\tau(Y^*(d_\xi)-a)\big],
\end{aligned}$$

where the first equality follows from the iterative expectation formula, the second inequality uses the no unmeasured confounder assumption. Note that $E\big[\rho_\tau(Y^*(d_\xi) - a)\big]$ is continuous and uniquely minimized by the τth quantile of $Y^*(d_\xi)$. Hence, the consistency follows from the standard M-estimation theory. $\square$

Recall that

$$\begin{aligned}
g(\cdot,\xi,m) &= \frac{I(C_\xi = \infty)}{\pi(\xi)} I(Y > m), \\
\widehat{m}_n &= \sup\{m : \sup_\xi P_n g(\cdot,\xi,m) \geq (1-\tau)\}, \\
m_0 &= \sup\{m : \sup_\xi P g(\cdot,\xi,m) \geq (1-\tau)\}, \\
\xi_0 &= \operatorname*{argmax}_\xi P g(\cdot,\xi,m_0).
\end{aligned}$$

and that we have $\widehat{\xi}_n = \operatorname*{argmax}_\xi P_n g(\cdot,\xi,\widehat{m}_n)$.

Lemma 2.2. *Under conditions (C1*)–(C4*), we have*

(1) $\widehat{m}_n = m_0 + O_p(n^{-1/2})$.

(2) $P_n g(\cdot,\widehat{\xi}_n, m_0) \geq \sup_\xi P_n g(\cdot,\xi,m_0) - o_p(n^{-2/3})$.

Proof. Part (1) follows from similar argument as that for Lemma 2(1) of the main paper. In the following, we derive part (2). By the standard M-estimation theory, $\widehat{\xi}_n$ is consistent for ξ_0. We will next show that $\widehat{\xi}_n - \xi_0 = O_p(n^{-1/3})$. Let $\theta = (\xi^T, \delta)^T$,

where $\delta = m - m_0$, and

$$h(\cdot, \xi, \delta) = \frac{I(C_\xi = \infty)}{\pi(\xi)} I(Y - m_0 - \delta > 0) - \frac{I(C_{\xi_0} = \infty)}{\pi(\xi_0)} I(Y - m_0 - \delta > 0).$$

By definition, $\widehat{\xi}_n = \underset{\xi}{\text{argmax}} P_n h(\cdot, \xi, \widehat{m}_n - m_0)$. We will consider a Taylor expansion of $Ph(\cdot, \xi, \delta)$ around $\theta_0 = (\xi_0^T, 0)^T$. Note that $h(\cdot, \xi_0, 0) = 0$ and that

$$\begin{aligned} & E\Big[\frac{I(C_\xi = \infty)}{\pi(\xi)} I(Y - m_0 - \delta > 0)\Big] \\ = & \sum_{(a_1,a_2)} E\Big[\frac{I(A_1 = d_\beta, A_2 = d_\gamma)}{P_1(H_1, \beta) P_2(H_2, \gamma)} I(Y - m_0 - \delta > 0) I(A_1 = a_1, A_2 = a_2)\Big], \end{aligned}$$

where (a_1, a_2) takes values in $\{(1,1), (1,0), (0,1), (0,0)\}$. For example, for $(a_1, a_2) = (1,1)$, the expectation is

$$\begin{aligned} & E\Big[\frac{I(H_1^T \beta > 0) I(H_2^T \gamma > 0)}{P_1(H_1, \beta) P_2(H_2, \gamma)} I(Y^*(1,1) - m_0 - \delta > 0) I(A_1 = 1, A_2 = 1)\Big] \\ = & \int\int I(h_1^T \beta > 0) I(h_2^T \gamma > 0) S(m_0 + \delta | h_2, a_2 = 1) f_{H_2|H_1,A_1}(h_2 | h_1, a_1 = 1) f_{H_1}(h_1) dh_2 dh_1, \end{aligned}$$

where $f_{H_1}(h_1)$ is the density function of H_1, $f_{H_2|H_1,A_1}(h_2|h_1, a_1)$ is the conditional density function of H_2 given $H_1 = h_1$ and $A_1 = a_1$, $S(\cdot|h_2, a_2)$ is the condition survival function of Y given the history $H_2 = h_2$ and $A_2 = a_2$. We obtain the above expectation formula because given the observed trajectory (X_1, A_1, X_2, A_2, Y), the likelihood is

$$f_{Y|H_2,A_2}(y|h_2, a_2) \pi_2(A_2 = a_2|h_2) f_{H_2|H_1,A_1}(h_2|h_1, a_1) \pi_1(A_1 = a_1|h_1) f_{H_1}(h_1),$$

where $f_{Y|H_2,A_2}(y|h_2, a_2)$ is the conditional density function of Y given $H_2 = h_2$ and

$A_2 = a_2$. Following similar calculations and letting

$$\begin{aligned}
R_1(h_1,\gamma,\delta) &= \int I(h_2^T\gamma > 0) S(m_0+\delta|h_2, a_2=1) f_{H_2|H_1,A_1}(h_2|h_1, a_1=1) dh_2,\\
R_2(h_1,\gamma,\delta) &= \int I(h_2^T\gamma \le 0) S(m_0+\delta|h_2, a_2=0) f_{H_2|H_1,A_1}(h_2|h_1, a_1=1) dh_2,\\
R_3(h_1,\gamma,\delta) &= \int I(h_2^T\gamma > 0) S(m_0+\delta|h_2, a_2=1) f_{H_2|H_1,A_1}(h_2|h_1, a_1=0) dh_2,\\
R_4(h_1,\gamma,\delta) &= \int I(h_2^T\gamma \le 0) S(m_0+\delta|h_2, a_2=0) f_{H_2|H_1,A_1}(h_2|h_1, a_1=0) dh_2,
\end{aligned}$$

we can write

$$E\big[h(\cdot,\xi,\delta)\big] = \sum_{k=1}^{4} G_k(\xi,\delta),$$

where for $j = 1, 2$,

$$G_j(\xi,\delta) = \int I(h_1^T\beta > 0) R_j(h_1,\gamma,\delta) f_{H_1}(h_1) dh_1 - \int I(h_1^T\beta_0 > 0) R_j(h_1,\gamma_0,\delta) f_{H_1}(h_1) dh_1;$$

and for $j = 3, 4$,

$$G_j(\xi,\delta) = \int I(h_1^T\beta \le 0) R_j(h_1,\gamma,\delta) f_{H_1}(h_1) dh_1 - \int I(h_1^T\beta_0 \le 0) R_j(h_1,\gamma_0,\delta) f_{H_1}(h_1) dh_1.$$

It is easy to see that $\frac{\partial}{\partial\delta} G_1(\xi,\delta)\big|_{\xi=\xi_0,\delta=0} = 0$ and $\frac{\partial^2}{\partial\delta^2} G_1(\xi,\delta)\big|_{\xi=\xi_0,\delta=0} = 0$. We have $\frac{\partial}{\partial\xi} G_1(\xi,\delta) = (\frac{\partial}{\partial\beta^T} G_1(\xi,\delta), \frac{\partial}{\partial\gamma^T} G_1(\xi,\delta))^T$. Similarly as the proof of Lemma 2(2) of the main paper, we consider the transformations $T_\beta = (I - ||\beta||^{-2}\beta\beta^T)(I - \beta_0\beta_0^T) + ||\beta||^{-1}\beta\beta_0^T$ and $T_\gamma = (I - ||\gamma||^{-2}\gamma\gamma^T)(I - \gamma_0\gamma_0^T) + ||\gamma||^{-1}\gamma\gamma_0^T$. We have

$$\begin{aligned}
&\frac{\partial}{\partial\beta^T} E(h(\cdot,\xi,\delta))\\
=\ & ||\beta||^{-2}\beta^T\beta_0(I + ||\beta||^{-2}\beta\beta^T) \int I\{h_1^T\beta_0 = 0\} R(T_\beta(h_1),\gamma,\delta) f_{H_1}(T_\beta(h_1)) h_1 d\sigma,
\end{aligned}$$

where $R(h_1, \gamma_0, \delta) = R_1(h_1, \gamma_0, \delta) + R_2(h_1, \gamma_0, \delta) - R_3(h_1, \gamma_0, \delta) - R_4(h_1, \gamma_0, \delta)$, σ is the surface measure on $\{H_1^T \beta_0 = 0\}$. Since $||\beta_0|| = 1$ and $T_{\beta_0}(h_1) = h_1$, we have

$$\frac{\partial}{\partial \beta^T} E(h(\cdot, \xi, \delta))\big|_{\xi=\xi_0, \delta=0} \;=\; (I + \beta_0 \beta_0^T) \int I\{h_1^T \beta_0 = 0\} R(h_1, \gamma_0, \delta) f_{H_1}(h_1) h_1 d\sigma.$$

Note that we have $\frac{\partial}{\partial \xi^T} E(h(\cdot, \xi, \delta))\big|_{\xi=\xi_0, \delta=0} = 0$ since $E(h(\cdot, \xi, 0))$ is maximized at $\xi = \xi_0$. This implies that

$$\int I\{h_1^T \beta_0 = 0\} R(h_1, \gamma_0, 0) f_{H_1}(h_1) h_1 d\sigma = 0.$$

Hence,

$$\begin{aligned} & \frac{\partial^2}{\partial \beta \partial \beta^T} E(h(\cdot, \xi, \delta))\big|_{\xi=\xi_0, \delta=0} \\ = \; & \int I\{h_1^T \beta_0 = 0\} \big[f_{H_1}(h_1) \dot{R}(h_1, \gamma, 0) + \dot{f}_{H_1}(h_1) R(h_1, \gamma, 0) \big]^T \beta_0 h_1 h_1^T d\sigma, \end{aligned}$$

where $\dot{f}_{H_1}(h_1)$ and $\dot{R}(h_1, \gamma, \delta)$ denote the derivatives with respect to h_1. We also have

$$\frac{\partial^2}{\partial \delta \partial \beta^T} E(h(\cdot, \xi, \delta))\big|_{\xi=\xi_0, \delta=0} \;=\; (I + \beta_0 \beta_0^T) \int I\{h_1^T \beta_0 = 0\} \frac{\partial}{\partial \delta} R(h_1, \gamma_0, 0) f_{H_1}(h_1) h_1 d\sigma,$$

where $\frac{\partial}{\partial \delta} R(h_1, \gamma_0, 0)$ is the partial derivative of $R(h_1, \gamma, \delta)$ with respect to δ evaluated at $(h_1, \gamma_0, 0)$. Now let

$$\begin{aligned} U_1(h_1, h_2, \delta) &= S(m_0 + \delta | h_2, a_2 = 1) f_{H_2|H_1, A_1}(h_2 | h_1, a_1 = 1) \\ &\quad - S(m_0 + \delta | h_2, a_2 = 0) f_{H_2|H_1, A_1}(h_2 | h_1, a_1 = 1) \\ U_2(h_1, h_2, \delta) &= S(m_0 + \delta | h_2, a_2 = 1) f_{H_2|H_1, A_1}(h_2 | h_1, a_1 = 0) \\ &\quad - S(m_0 + \delta | h_2, a_2 = 0) f_{H_2|H_1, A_1}(h_2 | h_1, a_1 = 0). \end{aligned}$$

We have

$$\begin{aligned}
&\frac{\partial}{\partial\gamma^T}E(h(\cdot,\xi,\delta))\\
=\;& r(\gamma)\int I\{h_1^T\beta>0\}f_{H_1}(h_1)\Big(\int I\{h_2^T\gamma_0=0\}U_1(h_1,T_\gamma(h_2),\delta)h_2d\sigma'\Big)dh_1\\
&+r(\gamma)\int I\{h_1^T\beta\le 0\}f_{H_1}(h_1)\Big(\int I\{h_2^T\gamma_0=0\}U_2(h_1,T_\gamma(h_2),\delta)h_2d\sigma'\Big)dh_1,
\end{aligned}$$

where $r(\gamma)=||\gamma||^{-2}\gamma^T\gamma_0(I+||\gamma||^{-2}\gamma\gamma^T)$, σ' is the surface measure on $\{H_2^T\gamma_0=0\}$. As in earlier calculation, we can show that

$$\begin{aligned}
&\int I\{h_1^T\beta_0>0\}f_{H_1}(h_1)\Big(\int I\{h_2^T\gamma_0=0\}U_1(h_1,h_2,0)h_2d\sigma'\Big)dh_1\\
&+\int I\{h_1^T\beta_0\le 0\}f_{H_1}(h_1)\Big(\int I\{h_2^T\gamma_0=0\}U_2(h_1,h_2,0)h_2d\sigma'\Big)dh_1=0.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\frac{\partial^2}{\partial\gamma\partial\gamma^T}E(h(\cdot,\xi,\delta))\big|_{\xi=\xi_0,\delta=0}\\
=\;&\int I\{h_1^T\beta_0>0\}f_{H_1}(h_1)\Big(\int I\{h_2^T\gamma_0=0\}\dot U_1(h_1,h_2,0)^T\gamma_0h_2h_2^Td\sigma'\Big)dh_1\\
&+\int I\{h_1^T\beta_0\le 0\}f_{H_1}(h_1)\Big(\int I\{h_2^T\gamma_0=0\}\dot U_2(h_1,h_2,0)^T\gamma_0h_2h_2^Td\sigma'\Big)dh_1,
\end{aligned}$$

where $\dot U_j(h_1,h_2,\delta)$ denotes the derivative with respect to h_2, $j=1,2$;

$$\begin{aligned}
&\frac{\partial^2}{\partial\beta\partial\gamma^T}E(h(\cdot,\xi,\delta))\big|_{\xi=\xi_0,\delta=0}\\
=\;&\int I\{h_1^T\beta_0=0\}f_{H_1}(h_1)\Big(\int I\{h_2^T\gamma_0=0\}\big(U_1(h_1,h_2,0)-U_2(h_1,h_2,0)\big)h_2d\sigma'\Big)h_1d\sigma;
\end{aligned}$$

and

$$\begin{aligned}
& \frac{\partial^2}{\partial\delta\partial\gamma^T}E(h(\cdot,\xi,\delta))\big|_{\xi=\xi_0,\delta=0} \\
= \;& (I+\gamma_0\gamma_0^T)\int I\{h_1^T\beta_0>0\}f_{H_1}(h_1)\Big(\int I\{h_2^T\gamma_0=0\}\frac{\partial}{\partial\delta}U_1(h_1,h_2,0)h_2 d\sigma'\Big)dh_1 \\
& +(I+\gamma_0\gamma_0^T)\int I\{h_1^T\beta_0\le 0\}f_{H_1}(h_1)\Big(\int I\{h_2^T\gamma_0=0\}\frac{\partial}{\partial\delta}U_2(h_1,h_2,0)h_2 d\sigma'\Big)dh_1,
\end{aligned}$$

where $\frac{\partial}{\partial\delta}U_j(h_1,h_2,0)$ denotes the partial derivative of $U_j(h_1,h_2,\delta)$ with respect to δ evaluated at (h_1,h_2,δ), $j=1,2$. Let

$$V^* = -\frac{\partial^2}{\partial\xi\partial\xi^T}E(h(\cdot,\xi,\delta))\big|_{\xi=\xi_0,\delta=0},$$

and let $a_1 = \frac{\partial^2}{\partial\xi^T\partial\delta}E(h(\cdot,\xi,\delta))\big|_{\xi=\xi_0,\delta=0}$. The Taylor expansion of $Ph(\cdot,\xi,\delta)$ around $(\xi_0,0)$ has the form

$$Ph(\cdot,\xi,\delta) \;=\; -\frac{1}{2}(\xi-\xi_0)^T V^*(\xi-\xi_0) + a_1^T(\xi-\xi_0)\delta + o(||\xi-\xi_0||^2) + o(\delta^2). \quad (1)$$

Let $H_R = \sup_{||\xi-\xi_0||\le R}|h(\cdot,\xi,\delta)|$. Note that for any given $0<R<1$, there exist positive constants b_1 and b_2 such that

$$\begin{aligned}
H_R &= \sup_{||\xi-\xi_0||\le R}\Big|\frac{I(C_\xi=\infty)}{\pi(\xi)}I(Y-m_0-\delta>0) - \frac{I(C_{\xi_0}=\infty)}{\pi(\xi_0)}I(Y-m_0-\delta>0)\Big| \\
&\le \sup_{||\xi-\xi_0||\le R}\Big|\frac{I(C_\xi=\infty)}{\pi(\xi)} - \frac{I(C_{\xi_0}=\infty)}{\pi(\xi_0)}\Big| + \sup_{||\xi-\xi_0||\le R}\Big|\frac{I(C_\xi=\infty)-I(C_{\xi_0}=\infty)}{\pi(\xi_0)}\Big| \\
&\le b_1||\xi-\xi_0|| + b_2\sup_{||\xi-\xi_0||\le R}\big|I(C_\xi=\infty)-I(C_{\xi_0}=\infty)\big| \\
&\le b_1||\xi-\xi_0|| + b_2\sup_{||\xi-\xi_0||\le R}\big|I(H_1^T\beta>0\ge H_1^T\beta_0)+I(H_1^T\beta_0>0\ge H_1^T\beta) \\
&\quad +I(H_2^T\gamma>0\ge H_2^T\gamma_0)+I(H_2^T\gamma_0>0\ge H_2^T\gamma)\big|.
\end{aligned}$$

Hence we have $E(H_R^2)=O(R)$ for R near zero. Hence, by Lemma 4.1 of Kim and

Pollard (1990), for each fixed $\epsilon > 0$, uniformly for $||\xi - \xi_0|| \leq R$,

$$P_n h(\cdot, \xi, \delta) \leq Ph(\cdot, \xi, \delta) + \epsilon(||\xi - \xi_0||^2 + \delta^2) + O_p(n^{-2/3}). \tag{2}$$

Combining this with (1), we have

$$P_n h(\cdot, \xi, \delta) \leq -\big(\frac{1}{2}\lambda_{min}(V^*) - \epsilon\big)||\xi - \xi_0||^2 + ||a_1||||\xi - \xi_0|||\delta| + (\epsilon + o(1))\delta^2 + O_p(n^{-2/3}).$$

Choosing $\epsilon = \lambda_{min}(V^*)/4$, we derive that

$$\begin{aligned} 0 &= P_n h(\cdot, \xi_0, \widehat{m}_n - m_0) \leq P_n h(\cdot, \widehat{\xi}_n, \widehat{m}_n - m_0) \\ &\leq -\frac{1}{4}\lambda_{min}(V^*)||\widehat{\xi}_n - \xi_0||^2 + O_p(n^{-1/2})||\widehat{\xi}_n - \xi_0|| + O_p(n^{-2/3}). \end{aligned}$$

Completing the square in $||\widehat{\xi}_n - \xi_0||$, we derive that $||\widehat{\xi}_n - \xi_0|| = O_p(n^{-1/3})$.

Next, we show that $\widehat{\xi}_n$ nearly maximizes $P_n h(\cdot, \xi, 0)$. A similar argument as above shows that $P|h(\cdot, \xi_1) - h(\cdot, \xi_2)| = O(||\xi_1 - \xi_2||)$ for ξ_1, ξ_2 near ξ_0. It follows from Lemma 4.6 of Kim and Pollard (1990) that the process

$$J_n(\cdot, \alpha, \eta) = n^{2/3}(P_n - P)h(\cdot, \xi_0 + \alpha n^{-1/3}, \eta n^{-1/3})$$

satisfies the stochastic equicontinuity condition of Theorem 2.3 of Kim and Pollard (1990). Since $n^{1/3}(\widehat{m}_n - m_0) = o_p(1)$, this implies, uniformly over $\xi - \xi_0$ in a $O_p(n^{-1/3})$ neighborhood of zero,

$$J_n(\cdot, n^{1/3}(\xi - \xi_0), n^{1/3}(\widehat{m}_n - m_0)) - J_n(\cdot, n^{1/3}(\xi - \xi_0), 0) = o_p(1).$$

That is,

$$P_n h(\cdot, \xi, \widehat{m}_n - m_0) = P_n h(\cdot, \xi, 0) + Ph(\cdot, \xi, \widehat{m}_n - m_0) - Ph(\cdot, \xi, 0) + o_p(n^{-2/3}),$$

uniformly over an $O_p(n^{-1/3})$ neighborhood of ξ_0. Within such a neighborhood, Taylor expansion similarly as before shows that $Ph(\cdot, \xi, \widehat{m}_n - m_0) - Ph(\cdot, \xi, 0) = o_p(n^{-2/3})$. Suppose $\widetilde{\xi}_n = \underset{\xi}{\text{argmax}}\ P_n h(\cdot, \xi, 0)$. An analysis similar to that for $\widehat{\xi}_n$ shows that $\widetilde{\xi}_n = \xi_0 + O_p(n^{-1/3})$. Hence,

$$\begin{aligned} P_n h(\cdot, \widehat{\xi}_n, 0) &= P_n h(\cdot, \widehat{\xi}_n, \widehat{m}_n - m_0) - o_p(n^{-2/3}) \\ &\geq P_n h(\cdot, \widetilde{\xi}_n, \widehat{m}_n - m_0) - o_p(n^{-2/3}) \\ &= P_n h(\cdot, \widetilde{\xi}_n, 0) - o_p(n^{-2/3}), \end{aligned}$$

where the inequality follows because $\widehat{\xi}_n = \underset{\xi}{\text{argmax}}\ P_n h(\cdot, \xi, \widehat{m}_n - m_0)$. Therefore,

$$P_n h(\cdot, \widehat{\xi}_n, 0) \geq \sup_{\xi} P_n h(\cdot, \xi, 0) - o_p(n^{-2/3}).$$

□

Proof of Theorem 2 of the main paper. Similarly as the proof of Theorem 1 of the main paper, it suffices to apply the main theorem of Kim and Pollard (1990) to the one parameter process $\{h(\cdot, \xi, 0) : \xi\}$. Recall that $h(\cdot, \xi, 0) = \frac{I(C_\xi = \infty)}{\pi(\xi)} I\{Y > m_0\} - \frac{I(C_{\xi_0} = \infty)}{\pi(\xi_0)} I\{Y > m_0\}$. In the following, we will verify conditions (iv) and (v) of the main theorem of Kim and Pollard (1990). For condition (iv), it can be shown that $\frac{\partial^2}{\partial \xi \partial \xi^T} E(h(\cdot, \xi, 0))\big|_{\xi = \xi_0} = -V^*$. Next, we calculate the kernel function in condition (v). To do so, for each C_1, C_2 in R^d, we will derive the limit of $tP\big|h(\cdot, \xi_0 + C_1/t, 0) -$

$h(\cdot, \xi_0 + C_2/t, 0)\big|^2$ as $t \to \infty$. We have

$$
\begin{aligned}
& tP\big|h(\cdot, \xi_0 + C_1/t, 0) - h(\cdot, \xi_0 + C_2/t, 0)\big|^2 \\
= \;& tP\Big[\Big(\frac{I(C_{\xi_0+C_1/t} = \infty)}{\pi(\xi_0 + C_1/t)} - \frac{I(C_{\xi_0+C_2/t} = \infty)}{\pi(\xi_0 + C_2/t)}\Big)^2 I(Y > m_0)\Big] \\
= \;& tP\Big[\frac{1}{\pi^2(\xi_0 + C_1/t)}\big(I(C_{\xi_0+C_1/t} = \infty) - I(C_{\xi_0+C_2/t} = \infty)\big)^2 I(Y > m_0)\Big] \\
& + tP\Big[I(C_{\xi_0+C_2/t} = \infty)\big(\frac{1}{\pi(\xi_0 + C_1/t)} - \frac{1}{\pi(\xi_0 + C_2/t)}\big)^2 I(Y > m_0)\Big] \\
& + 2tP\Big[\frac{1}{\pi(\xi_0 + C_1/t)}\big(I(C_{\xi_0+C_1/t} = \infty) - I(C_{\xi_0+C_2/t} = \infty)\big) I(C_{\xi_0+C_2/t} = \infty) \\
& \qquad \big(\frac{1}{\pi(\xi_0 + C_1/t)} - \frac{1}{\pi(\xi_0 + C_2/t)}\big) I(Y > m_0)\Big] \\
= \;& V_1 + V_2 + V_3.
\end{aligned}
$$

We will first evaluate V_1. We write $C_1 = (C_{11}^T, C_{12}^T)^T$, $C_2 = (C_{21}^T, C_{22}^T)^T$. Then

$$
\begin{aligned}
V_1 &= tP\Big\{\frac{1}{\pi^2(\xi_0 + C_1/t)}\Big[I(A_1 = d_{\beta_0+C_{11}/t}) I(A_2 = d_{\gamma_0+C_{12}/t}) \\
&\qquad - I(A_1 = d_{\beta_0+C_{21}/t}) I(A_2 = d_{\gamma_0+C_{22}/t})\Big]^2 I(Y > m_0)\Big\} \\
&= \sum_{j=1}^{4} V_{1j},
\end{aligned}
$$

where

$$
\begin{aligned}
V_{11} &= \int\int \frac{\pi(\xi_0)}{\pi^2(\xi_0 + C_1/t)} \Big[I(h_1^T(\beta_0 + C_{11}/t) > 0) I(h_2^T(\gamma_0 + C_{12}/t) > 0) \\
&\quad - I(h_1^T(\beta_0 + C_{21}/t) > 0) I(h_2^T(\gamma_0 + C_{22}/t) > 0) \Big]^2 S(m_0|h_2, a_2 = 1) \\
&\quad f_{H_2|H_1,A_1}(h_2|h_1, a_1 = 1) dh_2 f_{H_1}(h_1) dh_1, \\
V_{12} &= \int\int \frac{\pi(\xi_0)}{\pi^2(\xi_0 + C_1/t)} \Big[I(h_1^T(\beta_0 + C_{11}/t) > 0) I(h_2^T(\gamma_0 + C_{12}/t) \le 0) \\
&\quad - I(h_1^T(\beta_0 + C_{21}/t) > 0) I(h_2^T(\gamma_0 + C_{22}/t) \le 0) \Big]^2 S(m_0|h_2, a_2 = 0) \\
&\quad f_{H_2|H_1,A_1}(h_2|h_1, a_1 = 1) dh_2 f_{H_1}(h_1) dh_1, \\
V_{13} &= \int\int \frac{\pi(\xi_0)}{\pi^2(\xi_0 + C_1/t)} \Big[I(h_1^T(\beta_0 + C_{11}/t) \le 0) I(h_2^T(\gamma_0 + C_{12}/t) > 0) \\
&\quad - I(h_1^T(\beta_0 + C_{21}/t) \le 0) I(h_2^T(\gamma_0 + C_{22}/t) > 0) \Big]^2 S(m_0|h_2, a_2 = 1) \\
&\quad f_{H_2|H_1,A_1}(h_2|h_1, a_1 = 0) dh_2 f_{H_1}(h_1) dh_1, \\
V_{14} &= \int\int \frac{\pi(\xi_0)}{\pi^2(\xi_0 + C_1/t)} \Big[I(h_1^T(\beta_0 + C_{11}/t) \le 0) I(h_2^T(\gamma_0 + C_{12}/t) \le 0) \\
&\quad - I(h_1^T(\beta_0 + C_{21}/t) \le 0) I(h_2^T(\gamma_0 + C_{22}/t) \le 0) \Big]^2 S(m_0|h_2, a_2 = 0) \\
&\quad f_{H_2|H_1,A_1}(h_2|h_1, a_1 = 0) dh_2 f_{H_1}(h_1) dh_1.
\end{aligned}
$$

We first analyze V_{11}. Since

$$
\begin{aligned}
&I(h_1^T(\beta_0 + C_{11}/t) > 0) I(h_2^T(\gamma_0 + C_{12}/t) > 0) \\
&- I(h_1^T(\beta_0 + C_{21}/t) > 0) I(h_2^T(\gamma_0 + C_{22}/t) > 0) \\
= \; & I(h_1^T(\beta_0 + C_{11}/t) > 0) \big[I(h_2^T(\gamma_0 + C_{12}/t) > 0) - I(h_2^T(\gamma_0 + C_{22}/t) > 0) \big] \\
&+ I(h_2^T(\gamma_0 + C_{22}/t) > 0) \big[I(h_1^T(\beta_0 + C_{11}/t) > 0) - I(h_1^T(\beta_0 + C_{11}/t) > 0) \big],
\end{aligned}
$$

we can write $V_{11} = \sum_{j=1}^{3} V_{11j}$, where

$$
\begin{aligned}
V_{111} &= \int\int \frac{\pi(\xi_0)}{\pi^2(\xi_0 + C_1/t)} I(h_1^T(\beta_0 + C_{11}/t) > 0) \Big| I(h_2^T(\gamma_0 + C_{12}/t) > 0) \\
&\quad - I(h_2^T(\gamma_0 + C_{22}/t) > 0) \Big| S(m_0|h_2, a_2 = 1) f_{H_2|H_1,A_1}(h_2|h_1, a_1 = 1) dh_2 f_{H_1}(h_1) dh_1, \\
V_{112} &= \int\int \frac{\pi(\xi_0)}{\pi^2(\xi_0 + C_1/t)} I(h_2^T(\gamma_0 + C_{22}/t) > 0) \Big| I(h_1^T(\beta_0 + C_{11}/t) > 0) \\
&\quad - I(h_1^T(\beta_0 + C_{11}/t) > 0) \Big| S(m_0|h_2, a_2 = 1) f_{H_2|H_1,A_1}(h_2|h_1, a_1 = 1) dh_2 f_{H_1}(h_1) dh_1, \\
V_{113} &= \int\int \frac{\pi(\xi_0)}{\pi^2(\xi_0 + C_1/t)} I(h_1^T(\beta_0 + C_{11}/t) > 0) I(h_2^T(\gamma_0 + C_{22}/t) > 0) \\
&\quad \big[I(h_2^T(\gamma_0 + C_{12}/t) > 0) - I(h_2^T(\gamma_0 + C_{22}/t) > 0) \big] \\
&\quad \big[I(h_1^T(\beta_0 + C_{11}/t) > 0) - I(h_1^T(\beta_0 + C_{11}/t) > 0) \big] \\
&\quad S(m_0|h_2, a_2 = 1) f_{H_2|H_1,A_1}(h_2|h_1, a_1 = 1) dh_2 f_{H_1}(h_1) dh_1.
\end{aligned}
$$

Let $T_{\beta_0} = I - \beta_0 \beta_0^T$, $T_{\gamma_0} = I - \gamma_0 \gamma_0^T$. For any β, we can write $\beta = (\beta_0^T \beta)\beta_0 + T_{\beta_0}\beta$. Note that $T_{\beta_0}\beta$ is orthogonal to β_0 (that is $\beta_0^T T_{\beta_0}\beta = 0$). So we can decompose β as the sum of two components with the second component being orthogonal to the first. $\forall\ C_{11}$, we can write

$$\beta_0 + C_{11}/t = (1 + C_{11}^T \beta_0/t)\beta_0 + T_{\beta_0} C_{11}/t.$$

Similarly, we can write $h_1 = r_1\beta_0 + Z_1$, where Z_1 is orthogonal to β_0; $h_2 = r_2\beta_0 + Z_2$, where Z_2 is orthogonal to γ_0. Hence,

$$
\begin{aligned}
h_1^T(\beta_0 + C_{11}/t) &= (r_1\beta_0 + Z_1)^T \big[(1 + C_{11}^T\beta_0/t)\beta_0 + T_{\beta_0} C_{11}/t) \big] \\
&= r_1(1 + C_{11}^T \beta_0/t) + Z_1^T T_{\beta_0} C_{11}/t.
\end{aligned}
$$

Let $p_1(\cdot, \cdot|h_1, a_1)$ be the joint density function of (r_1, Z_1), and $p_2(\cdot, \cdot|h_1, a_1)$ be the joint conditional density function of (r_2, Z_2) given $H_1 = h_1$ and $A_1 = a_1$. We write

$C_{11}^* = T_{\beta_0} C_{11}$, $C_{12}^* = T_{\gamma_0} C_{12}$, $C_{21}^* = T_{\beta_0} C_{21}$ and $C_{22}^* = T_{\gamma_0} C_{22}$. Consider changes of variables $W_1 = tr_1$, $W_2 = tr_2$, we have

$$
\begin{aligned}
V_{111} \;=\; & \int \frac{\pi(\xi_0)}{\pi^2(\xi_0 + C_1/t)} I(h_1^T(\beta_0 + C_{11}/t) > 0) \\
& \Big[\int\int I\big(-(1 + C_{22}^T\gamma_0/t)^{-1} z_2^T C_{22}^* \geq w_2 > -(1 + C_{12}^T\gamma_0/t)^{-1} z_2^T C_{12}^* \big) \\
& S(m_0 \big| w_2\gamma_0/t + z_2, a_2 = 1) p_2(w_2/t, z_2 | h_1, a_1 = 1) dw_2 dz_2 \\
& + \int\int I\big(-(1 + C_{12}^T\gamma_0/t)^{-1} z_2^T C_{12}^* \geq w_2 > -(1 + C_{22}^T\gamma_0/t)^{-1} z_2^T C_{22}^* \big) \\
& S(m_0 \big| w_2\gamma_0/t + z_2, a_2 = 1) p_2(w_2/t, z_2 | h_1, a_1 = 1) dw_2 dz_2 \Big] f_{H_1}(h_1) dh_1, \\
\to \; & \int \frac{1}{\pi(\xi_0)} I(h_1^T\beta_0 > 0) \int \big| z_2^T (C_{12}^* - C_{22}^*) \big| S(m_0 \big| z_2, a_2 = 1) \\
& p_2(0, z_2 | h_1, a_1 = 1) dz_2 f_{H_1}(h_1) dh_1 \\
= \; & \int \frac{1}{\pi(\xi_0)} I(h_1^T\beta_0 > 0) \int \big| z_2^T (C_{12} - C_{22}) \big| S(m_0 \big| z_2, a_2 = 1) \\
& p_2(0, z_2 | h_1, a_1 = 1) dz_2 f_{H_1}(h_1) dh_1,
\end{aligned}
$$

when we integrate over w_2 and let $t \to \infty$, where the last equality follows because Z_2 is orthogonal to γ_0. Similar calculation yields

$$
\begin{aligned}
& V_{112} \\
\to \; & \int \frac{1}{\pi(\xi_0)} \big| z_1^T (C_{11} - C_{21}) \big| p_1(0, z_1) \int I(h_2^T\gamma_0 \geq 0) S(m_0 \big| z_2, a_2 = 1) \\
& f_{H_2|H_1,A_1}(h_2 | h_1, a_1 = 1) dh_2 dz_1,
\end{aligned}
$$

and $V_{113} \to 0$ as $t \to \infty$. Hence, we obtain the limit of V_{11} as $t \to \infty$. The limit of V_{1j}, $j = 2, 3, 4$, can be calculated similarly and yields the limit of V_1. We note that for t sufficient;y large, there exists a positive constant b_3 such that $|V_2| \leq b_3 tP\Big[\big(\pi(\xi_0 + C_1/t) - \pi(\xi_0 + C_2/t)\big)^2\Big] = O(t^{-1})$, which goes to 0 as $t \to \infty$. It is also straightforward to see $V_3 \to 0$ in probability as $t \to \infty$. Denote $\lim_{t\to\infty} tP|h(\cdot, \xi_0 + C_1/t, 0) - h(\cdot, \xi_0 + C_2/t, 0)|^2$ as $L(C_1 - C_2)$. Then similarly as the proof of Theorem 1, the limiting

covariance kernel function can be written as

$$K^*(C_1, C_2) = \frac{1}{2}(L(C_1) + L(C_2) - L(C_1 - C_2)).$$

□

2.3 Derivation of the theory for the optimal treatment regime minimizing Gini's mean difference criterion

Recall that $G(Y^*(d_\beta)) = -E\left|Y_1^*(d_\beta) - Y_2^*(d_\beta)\right|$, where $Y_1^*(d_\beta)$ and $Y_2^*(d_\beta)$ are independent copies of $Y^*(d_\beta)$. The optimal treatment regime that minimizes Gini's mean difference is defined as $\arg max_{d\in\mathbb{D}} G(Y^*(d_\beta))$. For a randomized study, we estimate $G(Y^*(d_\beta))$ using the second-order U-statistic $\widehat{G}(\beta) = -\frac{2}{n(n-1)}\sum_{1\le i<j\le n} 4C_i(\beta)C_j(\beta)|Y_i - Y_j|$. It follows from similar argument as that for Lemma 1 and the standard U statistic theory that for any given $\beta \in \mathbb{B}$, $\widehat{G}(\beta)$ is a consistent estimator of $G(Y^*(d_\beta))$.

The estimated parameter indexing the optimal rule is defined as $\widehat{\beta}_n = \underset{\beta\in\mathbb{B}}{\text{argmax}}\, \widehat{G}(\beta)$. Let $g(Z_i, Z_j, \xi, m) = 4C_i(\beta)C_j(\beta)(-|Y_i - Y_j| - m)$, $\widehat{m}_n = \sup\{m : \sup_{\beta\in\mathbb{B}} U_n g(\cdot,\cdot,\xi,m) \ge 0\}$, where $Z_i = \{X_i, Y_i\}$, $U_n g(\cdot,\cdot,\beta,m) = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} g(Z_i, Z_j, \xi, m)$. Then $\widehat{\beta}_n = \underset{\beta\in\mathbb{B}}{\text{argmax}}\, U_n g(\cdot,\cdot,\beta,\widehat{m}_n)$. Let $m_0 = \sup\{m : \sup_{\beta\in\mathbb{B}} Pg(\cdot,\cdot,\xi,m) \ge 0\}$ and $\beta_0 = \underset{\beta\in\mathbb{B}}{\text{argmax}}\, Pg(\cdot,\cdot,\xi,m_0)$.

The steps of deriving of the asymptotic distribution of $\widehat{\beta}_n$ are similar to those in the proof of Theorem 1. In the following, we highlight the new technical challenges in the derivation.

(1) To show $\widehat{m}_n = m_0 + O_p(n^{-1/2})$, we note that

$$\begin{aligned} & g(Z_i, Z_j, \beta, m) = 4C_i(\beta)C_j(\beta)(-|Y_i - Y_j| - m) \\ = \; & 4\big[A_i I(X_i^T\beta > 0) + (1 - A_i)I(X_i^T \leq 0)\big] \\ & \big[A_j I(X_j^T\beta > 0) + (1 - A_j)I(X_j^T \leq 0)\big] I(-|Y_i - Y_j| > m). \end{aligned}$$

The class $\{g(Z_i, Z_j, \beta, m) : \beta \in \mathbb{B}, m \in \mathbb{R}\}$ is a Euclidean class of functions with a constant envelope function. By Corollary 7 of Sherman (1994), we have

$$\sup_{\beta\in\mathbb{B}, m\in\mathbb{R}} \big|U_n g(\cdot,\cdot,\beta,m) - Pg(\cdot,\cdot,\beta,m)\big| = O_p(n^{-1/2}).$$

The rest of the proof follows from similar arguments as for Lemma 2(1).

(2) To show $U_n h(\cdot,\cdot,\widehat{\beta}_n, m_0) \geq \sup_{\beta\in\mathbb{B}} U_n h(\cdot,\beta,m_0) - o_p(n^{-2/3})$, we first note that the consistency of $\widehat{\beta}_n$ follows from standard M-estimation theory. To show $\widehat{\beta}_n - \beta_0 = O_p(n^{-1/3})$, we let

$$\begin{aligned} & h(Z_i, Z_j, \beta, \delta) \\ = \; & 4C_i(\beta)C_j(\beta)(-|Y_i - Y_j| - m_0 - \delta) - 4C_i(\beta_0)C_j(\beta_0)(-|Y_i - Y_j| - m_0 - \delta). \end{aligned}$$

By definition, $\widehat{\beta}_n = \underset{\beta\in\mathbb{B}}{\text{argmax}}\; U_n h(\cdot,\beta,\widehat{m}_n - m_0)$. We'll consider a Taylor expansion of $Ph(Z_i, Z_j, \beta, \delta)$ around $\theta_0 = (\beta_0^T, 0)^T$. Note that $h(Z_i, Z_j, \beta_0, 0) = 0$ and that for a

randomized trial

$$
\begin{aligned}
& E\big[4C_i(\beta)C_j(\beta)(-|Y_i - Y_j| - m_0 - \delta)\big] \\
= \; & E\big\{I(X_i^T\beta > 0)I(X_j^T\beta > 0)I(-|Y_i^*(1) - Y_j^*(1)| - m_0 - \delta > 0)\big\} \\
& + E\big\{I(X_i^T\beta > 0)I(X_j^T\beta \le 0)I(-|Y_i^*(1) - Y_j^*(0)| - m_0 - \delta > 0)\big\} \\
& + E\big\{I(X_i^T\beta \le 0)I(X_j^T\beta > 0)I(-|Y_i^*(0) - Y_j^*(1)| - m_0 - \delta > 0)\big\} \\
& + E\big\{I(X_i^T\beta \le 0)I(X_j^T\beta \le 0)I(-|Y_i^*(0) - Y_j^*(0)| - m_0 - \delta > 0)\big\} \\
= \; & E\big\{I(X_i^T\beta > 0)I(X_j^T\beta > 0)S_{11,X_i,X_j}(m_0 + \delta)\big\} \\
& + E\big\{I(X_i^T\beta > 0)I(X_j^T\beta \le 0)S_{10,X_i,X_j}(m_0 + \delta)\big\} \\
& + E\big\{I(X_i^T\beta \le 0)I(X_j^T\beta > 0)S_{01,X_i,X_j}(m_0 + \delta)\big\} \\
& + E\big\{I(X_i^T\beta \le 0)I(X_j^T\beta \le 0)S_{00,X_i,X_j}(m_0 + \delta)\big\},
\end{aligned}
$$

where S_{11,X_i,X_j} is the survival function of $-|Y_i^*(1) - Y_j^*(1)|$ given (X_i, X_j), S_{10,X_i,X_j} is the survival function of $-|Y_i^*(1) - Y_j^*(0)|$ given (X_i, X_j), S_{01,X_i,X_j} and S_{00,X_i,X_j} are defined similarly. To calculate the V matrix in the Taylor expansion, we need to evaluate the derivative (with respect to β) of the surface integral of the form

$$
\iint I(X_1^T\beta > 0, X_2^T\beta > 0)S_{11,X_1,X_2}(m_0 + \delta)f(X_1)f(X_2)dX_1dX_2.
$$

This can be done by applying the general formula of Uryasev (1995). Furthermore, we note that Lemma 4.1 of Kim and Pollard (1990) can be readily extended to U-statistics by applying the U-process maximal inequality of Sherman (1994). Hence, it follows similar argument as in the proof of Lemma 2(2) that $\widehat{\beta}_n - \beta_0 = O_p(n^{-1/3})$. Finally, we can show that $\widehat{\beta}_n$ nearly maximizes $U_n h(\cdot, \cdot, \beta, 0)$. We consider the centered U-process

$$
J_n(\cdot, \cdot, \alpha, \gamma) = n^{2/3}(U_n - P)h(\cdot, \cdot, \beta_0 + \alpha n^{-1/3}, \gamma n^{-1/3}).
$$

It can be shown that this process is stochastic equicontinous by applying the same argument as for Lemma 4.6 of Kim and Pollard (1990) and appealing to the U-process maximal inequality in Sherman (1994). This allows us to apply similar argument as that for Lemma 2(2) to show that $U_n h(\cdot,\cdot,\widehat{\beta}_n, m_0) \geq \sup_{\beta\in\mathbb{B}} U_n h(\cdot,\beta,m_0) - o_p(n^{-2/3})$.

(3) To derive the asymptotic distribution, we first note based on the result in (2), we only need to consider the one-parameter U process $U_n h(\cdot,\cdot,\beta,m_0)$. However, the main theorem of Kim and Pollard (1990) is still not applicable as it is stated for estimators minimizing a sample average (not U process). We consider the Hoeffding decomposition of $U_n h(\cdot,\cdot,\beta_n,m_0)$:

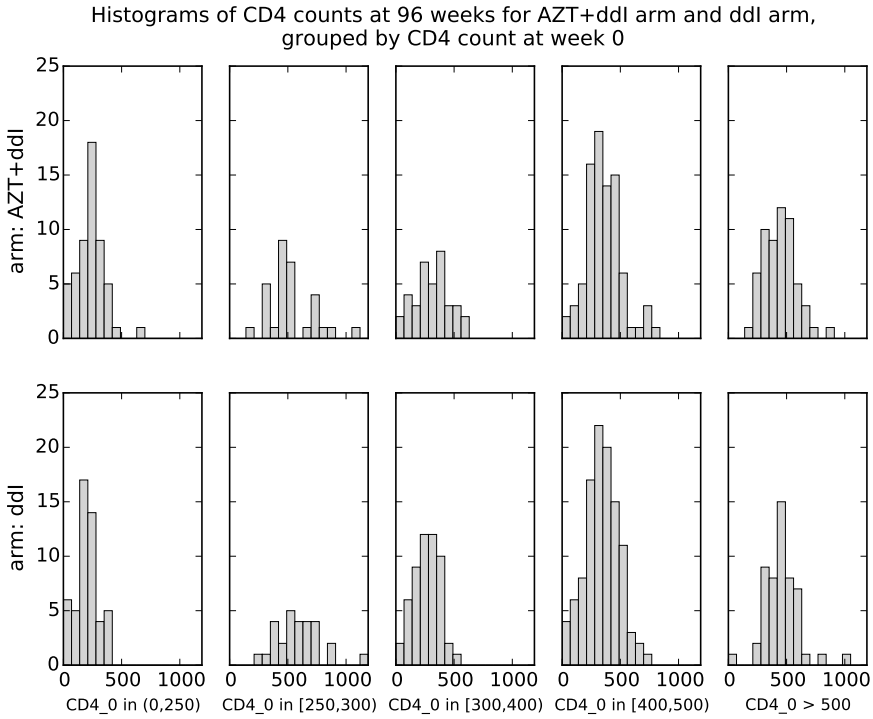
$$\begin{aligned} & U_n h(Z_i,Z_j,\beta,m_0) \\ = \;& P(h(Z_i,Z_j,\beta,m_0)) + \frac{2}{n}\sum_{i=1}^{n} h^*(Z_i,\beta,m_0) + U_n\tilde{h}(Z_i,Z_j,\beta,m_0), \end{aligned}$$

where

$$\begin{aligned} h^*(z,\beta,m_0) &= P\big[h(z,Z_j,\beta,m_0)\big] - P(h(\cdot,\cdot,\beta,m_0)), \\ \tilde{h}(Z_i,Z_j,\beta,m_0) &= h(Z_i,Z_j,\beta,m_0) - h^*(Z_i,\beta,m_0) - h^*(Z_j,\beta,m_0) + P(h(Z_i,Z_j,\beta,m_0)). \end{aligned}$$

Note that $U_n\tilde{h}(Z_i,Z_j,\beta,m_0)$ is a second-order degenerate U-process. Corollary 4 of Sherman (1994) implies that $\sup_{\beta\in\mathbb{B}}\left|U_n\tilde{h}(Z_i,Z_j,\beta,m_0)\right| = O(n^{-1})$. Hence, we can restrict to the leading terms of the Hoeffding decomposition. Let $\eta(Z_i,\beta,m_0) = 2h^*(Z_i,\beta,m_0)+P(h(Z_i,Z_j,\beta,m_0))$, then we have $P_n\eta(Z_i,\widehat{\beta}_n,m_0) \geq \sup_{\beta\in\mathbb{B}} P_n h(Z_i,\beta,m_0) - o_p(n^{-2/3})$. We then can finish the proof by similar arguments as in the proof of Theorem 1 by considering $\eta(Z_i,\beta,m_0)$. □

Figure 1

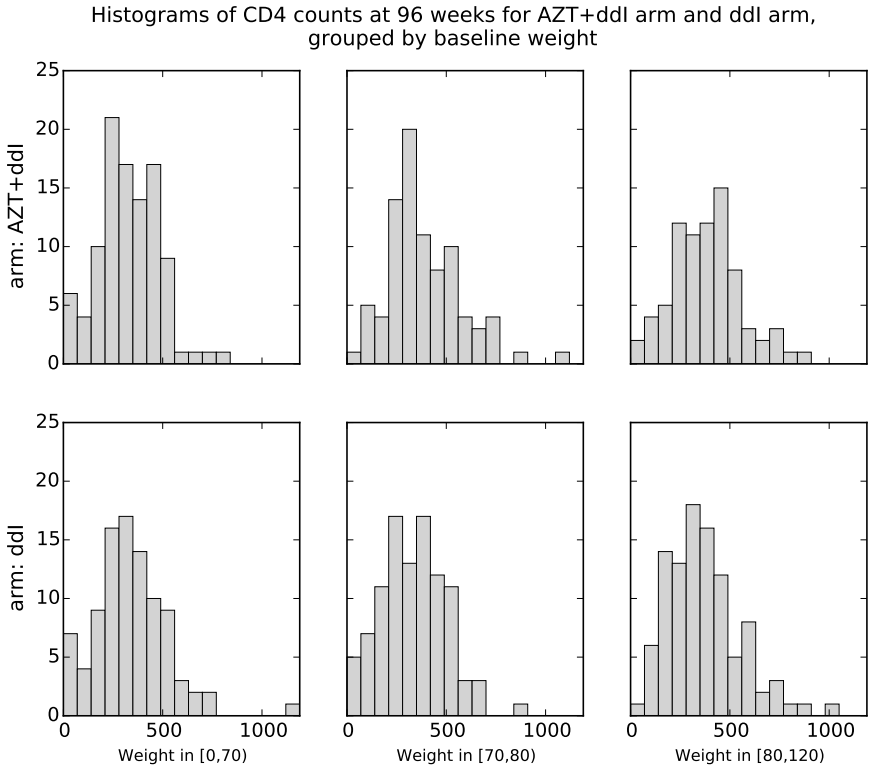


3 Additional numerical results

3.1 Additional graphs for ACTG175

Two important covariates in the real data example of ACTG175 study discussed in the main paper are CD4 count at week 0 and baseline weight. Figure 1 and Figure 2 display the histograms of the response variable (CD4 count at week 96) for each of the two treatment arms for different subgroups of patients, for which the subgroups are formed by the observed values of the CD4 count at week 0 or baseline weight. The varying shapes of the histograms across different ranges of both covariates indicate heteroscedastic treatment effects. It is also observed that the distribution of the response variable tends to be asymmetric and skewed to the right. Figure 3 depicts the three estimated regimes graphically, from which we observe that they are dramatically different from each other.

Figure 2



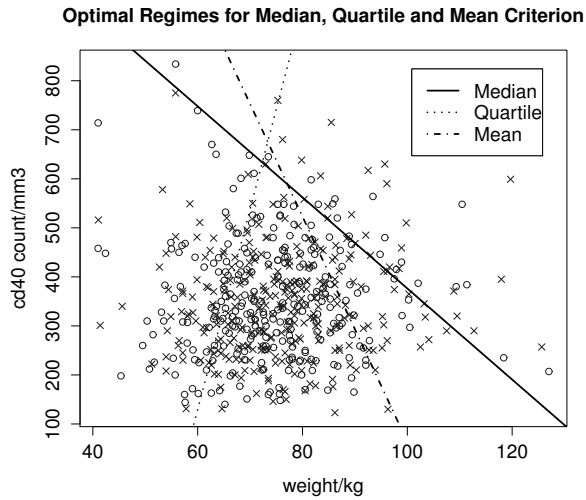


Figure 3: Graphical representation of the estimated optimal treatment regimes for ACTG175 data analysis.

Table 1: Proportions of optimal assignment evaluated by different criteria (with standard deviations in the parenthesis) for Example 1 of the main paper.

		Theoretically optimal regime		
Estimated optimal regime	n	mean criterion	0.25qt criterion	0.10qt criterion
mean criterion	500	0.84(0.06)	0.73(0.08)	0.57(0.09)
	1000	0.86(0.04)	0.74(0.07)	0.57(0.08)
	5000	0.92(0.02)	0.76(0.04)	0.57(0.05)
0.25qt criterion	500	0.74(0.07)	0.85(0.05)	0.73(0.08)
	1000	0.74(0.06)	0.88(0.04)	0.76(0.07)
	5000	0.75(0.03)	0.92(0.02)	0.77(0.04)
0.10qt criterion	500	0.58(0.08)	0.76(0.07)	0.88(0.04)
	1000	0.57(0.05)	0.75(0.05)	0.91(0.03)
	5000	0.57(0.03)	0.75(0.03)	0.94(0.01)

3.2 Additional simulation results for Example 1 of Section 5.1 of the main paper

Table 1 summarizes the proportion of times the estimated optimal treatment regime matches the theoretically optimal treatment regime across the sample by different criteria, for the three different sample sizes. The standard deviation of the proportion is reported in the parenthesis. More specially, for each simulation run, we apply the estimated optimal treatment regime to assign each subject in the sample to treatment 0 or 1, then we evaluate what proportion of these assignments are consistent with the assignments if a theoretically optimal criterion is applied. To save space, for the mean criterion, we only apply the model-free estimator mean_ZTLD. We observe that (1) the proportion of subjects who actually receive the theoretically optimal assignments when the estimated optimal treatment regime is applied as quite high and increases with the sample size; (2) if we assign the subjects using the estimated mean-optimal criterion, then a significant proportion of the assignments are suboptimal if evaluated by the quantile criterion; and vice versa.

3.3 A numerical example for the doubly robust estimator in Section 1.1 of this online supplement

We generate a random sample (Y_i, A_i, X_i) with $X_i = (X_{i1}, X_{i2})^T$ from a heteroscedastic regression model

$$\begin{aligned} Y_i &= X_{i1}^2 + X_{i2}^2 + A_i \exp(0.11 - X_{i1} - X_{i2}) + A_i \text{Gamma}(\text{shape} = 2X_{i1} + 3, \text{scale} = 1) \\ &\quad + (1 - A_i)\,\text{N}(\text{mean} = 2X_{i1} + 3, \text{sd} = 0.5), \end{aligned}$$

where X_{i1}, X_{i2} are independent and uniformly distributed on $[-1.5, 1.5]$, $A_i \big| X_i$ is Bernoulli with success probability satisfying $\text{logit}\{P(A_i = 1 \mid X_i)\} = 1 - X_{i1}^2 - X_{i2}^2$, $i = 1, \ldots, 1000$. We consider estimating the optimal treatment regime for maximizing the 0.3 quantile of the potential outcome distribution when the propensity score model may not be correctly specified. We consider the class of treatment regimes $I(\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0)$, where $(\beta_0, \beta_1, \beta_2)^T$ has L_2-norm 1. Based on a Monte Carlo experiment with sample size 10^5, the optimal treatment regime is given by $I(0.530 - 0.814X_1 - 0.230X_2 > 0)$ and the largest achievable 0.3 quantile of the potential outcome distribution over the class is 4.841.

To compare the original inverse probability weighted estimator in Section 3.1 of the main paper and the doubly robust estimator (DR) in this section, we consider two models for the propensity score function. One is the correctly specified model $\text{logit}\{\pi(X_i; \gamma)\} = \gamma_0 + \gamma_1 X_{i1}^2 + \gamma_2 X_{i2}^2$; the other is a misspecified model $\text{logit}\{\pi(X_i; \gamma)\} = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2}$. The results are summarized in Table 2. When the propensity score function is specified correctly, both the original estimator and the DR estimator have satisfactory performance. However, when the propensity score function is misspecified, the original estimator displays substantial bias while the DR estimator is still accurate.

Table 2: Doubly robust estimation for a heteroskedastic model: estimated parameters indexing the optimal treatment regimes and their corresponding estimated value functions.

Method	PS model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\widehat{Q}_{0.3}$
Original estimator	correct	0.524(0.08)	-0.782(0.085)	-0.242(0.204)	4.916(0.124)
	incorrect	0.054(0.264)	-0.915(0.051)	-0.265(0.134)	4.785(0.119)
DR estimator	correct	0.530(0.077)	-0.786(0.119)	-0.196(0.206)	4.901(0.127)
	incorrect	0.524(0.082)	-0.801(0.079)	-0.210(0.165)	4.921(0.136)

Table 3: Estimated optimal treatment regimes and estimated value functions under Gini's mean difference criterion ($n = 300$, 200 runs)

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\widehat{G}(\hat{\beta})$
mean of estimates	0.58	-0.56	-0.56	-1.57
std. error	(0.08)	(0.15)	(0.15)	(0.14)

3.4 A numerical example for Gini's mean difference criterion in Section 1.2 of this online supplement

We generate random observations $\{Y_i, A_i, X_i\}$ from

$$Y_i = \left[1 + 0.5\left(X_{i1} + X_{i2}\right) + A_i\left(X_{i1} + X_{i2} - 0.4\right)\right]^2 + \left(2 - 1.5A_i\right)\epsilon_i,$$

where $X_{ik} \sim \text{Uniform}\,(0,1)$ i.i.d., and $\epsilon_i \sim N(0,1)$ are mutually independently distributed. The treatment indicator A_i (1 for treatment, 0 for control) was generated according to a logistic model: $\text{logit}\left[P\left(A_i = 1|X_i\right)\right] = -0.5 + X_{i1}^2 + X_{i2}^2$.

We seek for the treatment regime d_β that maximizes $G(Y^*(d_\beta)) = -E\left|Y_1^*(d_\beta) - Y_2^*(d_\beta)\right|$. We consider the class of treatment regimes of the form $I\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1} > 0\right)$, where $\beta = (\beta_0, \beta_1, \beta_2)^T$ is normalized to have L_2 norm equal to 1. Using a simulated Monte Carlo data set of 10^5 observations, we obtain that the approximate true optimal treatment regime is indexed by $(0.5797, -0.5741, -0.5781)$, according to which the negative of the mean absolute difference of the distribution of the potential outcome is -1.855. The results of estimated regimes are presented in Table 3. The estimators are close to the population values.

Table 4: Population parameters and values for optimal treatment regimes under the weighted quantile criterion. The last four columns are calculated via a Monte Carlo data set with $n = 10^5$.

	β_0	β_1	β_2	$Q^{\mathbf{w}_1}(\beta)$	$Q^{\mathbf{w}_2}(\beta)$	$Q^{\mathbf{w}_3}(\beta)$	$Q^{\mathbf{w}_4}(\beta)$
Setting 1	0.39	-0.67	-0.62	0.74	0.98	1.49	1.63
Setting 2	0.44	-0.62	-0.65	0.73	0.99	1.52	1.67
Setting 3	0.54	-0.59	-0.60	0.59	0.92	1.57	1.73
Setting 4	0.57	-0.59	-0.57	0.44	0.84	1.56	1.74

3.5 A numerical example for the weighted quantile criterion in Section 1.2 of this online supplement

We generate random observations (Y_i, A_i, X_i), with $X_i = (X_{i1}, X_{i2})'$, such that

$$Y_i = 1 + X_{i1} + X_{i2} + A_i\left(3 - 2.5X_{i1} - 2.5X_{i2}\right) + \left(1 + A_i\left(1 + X_{i1} + X_{i2}\right)\right)\epsilon_i, \qquad (3)$$

where $X_{ik} \sim \text{Uniform}\,(0,1)$ i.i.d, and $\epsilon_i \sim N(0,1)$. The treatment indicator $A_i \in \{0,1\}$ was generated according to a logistic regression model: $\text{logit}\left(\Pr\left(A = 1|X\right)\right) = -0.5 + X_1 + X_2$.

In this example we consider $\boldsymbol{\tau} = (0.1, 0.3)$ and four different choices of weight vector: (1) $\mathbf{w}_1 = (1,0)$, (2) $\mathbf{w}_2 = (0.8, 0.2)$, (3) $\mathbf{w}_3 = (0.2, 0.8)$, and (4) $\mathbf{w}_4 = (0,1)$. It is easy to see setting 1 degenerates to 0.10 quantile criterion, and setting 4 degenerates to 0.30 quantile criterion. Table 4 summarizes the population parameters and values for optimal treatment regimes under the weighted quantile criterion, which are obtained via a Monte Carlo data set with 10^5 observations. The estimated optimal treatment regimes and estimated value functions are reported in Table 5 for sample size 1000 and 200 simulation runs. The estimated values are observed to be close to their targets.

Table 5: Estimated optimal treatment regimes and estimated value functions under the weighted quantile criterion.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\widehat{Q}^{\mathbf{w}_1}\left(\hat{\beta}\right)$	$\widehat{Q}^{\mathbf{w}_2}\left(\hat{\beta}\right)$	$\widehat{Q}^{\mathbf{w}_3}\left(\hat{\beta}\right)$	$\widehat{Q}^{\mathbf{w}_4}\left(\hat{\beta}\right)$
Setting 1	0.40(0.07)	-0.62(0.16)	-0.63(0.16)	0.82(0.10)	1.04(0.10)	1.54(0.09)	1.68(0.10)
Setting 2	0.44(0.06)	-0.61(0.15)	-0.62(0.15)	0.78(0.10)	1.07(0.10)	1.58(0.10)	1.73(0.11)
Setting 3	0.54(0.05)	-0.57(0.13)	-0.59(0.13)	0.58(0.20)	0.95(0.14)	1.65(0.10)	1.80(0.11)
Setting 4	0.57(0.06)	-0.56(0.13)	-0.57(0.13)	0.39(0.31)	0.82(0.23)	1.62(0.11)	1.82(0.11)

3.6 A numerical example investigating the effect of including more covariates in the treatment regime space

An anonymous referee suggested that "there is a trade-off regarding the number of covariates to be included in X. If X includes, say, only gender, we estimate only treatment effects for men and women and thus our statistical treatment choice rules will be very simple: One recommendation for men and one recommendation for women. On the other hand, if X includes very many regressors, the set of recommendations will be much richer, i.e. we have tailored recommendations for gender by age by race by education... Yet, these recommendations will be less precisely estimated and thus more likely to be wrong". This is an insightful comment and it is true for any existing method of estimating optimal treatment regime. We demonstrate this using a simple example. We generate random observations from the following model

$$Y_i = 1 + A_i - 1.25A_iX_{i1} + (2 - 1.8A_iX_{i1})\,\epsilon_i, \quad i = 1, \ldots, n, \tag{4}$$

where X_{i1} is binary and follows Bernoulli(0.5) distribution, $\epsilon_i \sim N(0,1)$ is independent of X_{i1}. We also generate $X_{i2} \sim N(0,1)$, independent of ϵ_i and X_{i1}. In our example, X_{i2} is a redundant covariate that is not related to the outcome variable.

We consider estimating median-optimal treatment regimes in the following two classes of treatment regimes: (1) $\mathbb{D}_1 = \{s_0 I\{X_1 = 0\} + s_1 I\{X_1 = 1\} | s_0, s_1 \in \{0,1\}\}$, and (2) $\mathbb{D}_2 = \Big\{ s_{00} I\{X_1 = 0, X_2 < 1\} + s_{01} I\{X_1 = 0, X_2 > 1\} + s_{10} I\{X_1 = 1, X_2 <$

$1\}+s_{11}I\{X_1=1,X_2>1\}|s_{00},s_{01},s_{10},s_{11}\in\{0,1\}\Big\}$. Class $\mathbb{D}_1$ has 4 distinct treatment regimes; class $\mathbb{D}_2$ has 2^4 distinct treatment regimes. Class $\mathbb{D}_2$ is an enrichment of $\mathbb{D}_1$ using the redundant covariate X_2. Evaluated using a large Monte Carlo sample, the median-optimal rule in class $\mathbb{D}_1$ corresponds to $s_0=1$ and $s_1=0$ (i.e, the treatment regime $I\{X_1=0\}$), which is also the median-optimal treatment regime in class $\mathbb{D}_2$ (i.e., $(s_{00},s_{01},s_{10},s_{11})=(1,1,0,0)$). This is regarded as the theoretically median-optimal treatment regime. The corresponding maximally achievable median value of the potential outcome is 1.501.

Based on a random sample of size 500, we estimate the median-optimal treatment regimes in each of the two classes and found that for class $\mathbb{D}_1$, the proportion of times the estimated median-optimal treatment regime is equivalent to the theoretically median-optimal treatment regime is 99.9% (with a standard deviation is 0.3%), with the estimated maximally achievable median value of the potential outcome equal to 1.499; for class $\mathbb{D}_2$, the proportion of times the estimated median-optimal treatment regime is equivalent to the theoretically median-optimal treatment regime is 83.4% (with a standard deviation is 3.1%), with the estimated maximally achievable median value of the potential outcome equal to 1.497.

Such kind of trade-off is expected to happen for any other methods in the literature for estimating optimal treatment regime. An important future research topic is thus variable selection for estimating optimal treatment regime.

3.7 Algorithm implementation

The methods proposed in this paper can be implemented using the R package *quantoptr* (Zhou et al., 2017). Following a referee's suggestion, we provide here the pseudocode for estimating the quantile-optimal treatment regime. The pseudocode in Listing 1 outlines how to estimate the marginal quantile for a treatment regime index

by β (beta) given the following input: design matrix x, observed outcomes y, observed treatment a, the estimated value of $P(A = 1 \mid X)$ $(prob)$ and the quantile level of interest (tau). In this pseudocode, g denotes the treatment regime indexed by β, c is the missingness indicator, *wts* is the vector of $\frac{C_i(\beta)}{\hat{\pi}_c(X_i,\beta)}$. The output of the pseudocode *quantile* is the estimator $\widehat{Q}_\tau(\beta)$, i.e., the estimated marginal quantile of the potential outcome when the treatment regime is indexed by β. The estimated marginal quantile is obtained using inverse probability weighting, which is implemented using a weighted quantile regression via the function "rq" in the R package quantreg (Koenker, 2016).

Listing 1: Pseudocode for estimating the marginal quantile of potential outcome

```
quant_est <- function(beta,x,y,a,prob,tau){
  g = I( matrix.multiply(x,beta) > 0)
  c = a*g+(1-a)*(1-g)
  wts = g*1/prob+(1-g)*(1/(1-prob))
  wts = c*wts
  quantile = weighted.quantile(y, wts, tau)
  return(quantile)
}
```

Next, to estimate the parameter indexing the optimal treatment regime, we adopt the genetic algorithm and use the function "genoud" in the R package rgenoud (Mebane and Sekhon, 2011), which was first recommended in Zhang et al. (2012) for optimizing a nonsmooth objective function in the setting of estimating mean-optimal treatment regime. Listing 2 provides the pseudocode that outlines how the "genoud" function was used to derive the optimal quantile treatment rule and corresponding estimated quantile value. The "genoud" function takes the function "quant_est" in the pseudocode in Listing 1 as part of the input to find an optimal treatment regime when β is allowed to vary in a given domain. The domain for each parameter is set to

be (-1,1) because $||\beta|| = 1$ and the starting value is the zero-vector. The "Nelder-Mead" optimization method (Nelder and Mead,1965) is used because the objective function is discontinuous. The output *beta_hat* is the estimated vector of parameters that indexes the quantile-optimal treatment regime, and the output *q_hat* is the estimated τth quantile of the potential outcome corresponding to the estimated optimal treatment regime.

Listing 2: Pseudocode for estimating the optimal treatment regime

```
beta_est<-function(x,y,a,prob,tau){
  genoud_results = genoud(function=quant_est,x=x,y=y,a=a,prob=prob,
                          tau=tau, Domains = (-1,1),
                          starting.values=0,
                          optim.method="Nelder-Mead")
  beta_hat = optimal.parameter(genoud_results)
  q_hat = optimal.value(hatQ)
  return(beta_hat,q_hat)
}
```

References

[1] Gini, C. (1912). Variabilitàe mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. *Studi Economico-giuridici della Regia Facoltà Giurisprudenza*, parte II, Cuppini, Bologna.

[2] Kim, J. K. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, **1**, 191-219.

[3] Koenker, R. (2005). *Quantile Regression*. Cambridge University Press

[4] Koenker, R. (2016). quantreg: Quantile regression. https://cran.r-project.org/web/ packages/quantreg/index.html.

[5] Mebane Jr, W.R. and Sekhon, J.S. (2011). Genetic optimization using derivatives: the rgenoud package for R. *Journal of Statistical Software*, 42(11), 1-26.

[6] Nelder, J. and Mead, R. (1965). A simplex algorithm for function minimization. Computer Journal, 7(4):308-313.

[7] Sherman, R. P. (1994). Maximal inequalities for degenerate U-processes with applications to optimization estimators. *The Annals of Statistics*, 439-459.

[8] Uryasev, S. (1995). Derivatives of probability functions and some applications. *Annals of Operations Research*, **56**, 287-311.

[9] Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010-1018.

[10] Zhou, Y., Wang, L., Sherwood, B. and Song, R. (2017) quantoptr: Algorithms for Quantile- And Mean-Optimal Treatment Regimes, https://CRAN.R-project.org/package=quantoptr